The very model of a modern linguist

—

in honor of Helge Dyvik

BeLLS Volume 8 (2017)

Edited by
Victoria Rosén and Koenraad De Smedt

Bergen Language and Linguistics Studies
ISSN 1892-2449
ISBN 978-82-93643-01-2 (electronic) • 978-82-93643-00-5 (print)
http://dx.doi.org/10.15845/bells.v8i1

Table of Contents

Tabula Gratulatoria v

Contributors ix

Preface xv

Old English and Old Norwegian noun phrases with two attributive adjectives . 1
Kristin Bech

Judgement, taste and closely related Germanic languages 19
Robin Cooper

Unlike phrase structure category coordination 33
Mary Dalrymple

Finite-state tokenization for a deep Wolof LFG grammar 56
Cheikh M. Bamba Dione

Syntactic discontinuities in Latin – A treebank-based study 75
Dag Haug

Increasing grammar coverage through fine-grained lexical distinctions 97
Petter Haugereid

A word or two? 112
Christer Johansson

Preserving grammatical functions in LFG 127
Ronald M. Kaplan

Norwegian *masse*: from measure noun to quantifier 143
Torodd Kinn

Reflexive sentences with *la* 'let' in Norwegian — active or passive? 167
Helge Lødrup

From LFG structures to dependency relations 183
Paul Meurer

A full-fledged hierarchical lexicon in LFG: the FrameNet approach 202
Adam Przepiórkowski

Norwegian bare singulars revisited 220
Victoria Rosén & Kaja Borthen

The concept of 'translation unit' revisited 241
Martha Thunes

Subject properties in presentational sentences in Icelandic and Swedish . . . 260
Annie Zaenen, Elisabet Engdahl & Joan Maling

Tabula Gratulatoria

Jardar Eggesbø Abrahamsen, Trondheim
Lars Ahrenberg, Linköping
Gunnstein Akselberg, Bergen
Gulbrand Alhaug, Tromsø
Jens Allwood, Gothenburg
Nazareth Amlesom Kifle, Halden
Øivin Andersen, Bergen
Gisle Andersen, Bergen
Erik Andvik, Bergen
Marianna Apidianaki, Paris
John Ole Askedal, Oslo
Jon Askeland, Bergen
Jóhanna Barðdal, Ghent
Kristin Bech, Oslo
Bergljot Behrens, Oslo
Martin van den Berg, Palo Alto
Harald Berggreen, Bergen
Magne Bergland, Bergen
Julie Bergmann, Bergen
Eckhard Bick, Aarhus
Chris Biemann, Hamburg
Kristín Bjarnadóttir, Reykjavík
Anne Kari Bjørge, Bergen
Tove Bjørneset, Bergen
Francis Bond, Singapore
Lars Borin, Gothenburg
Kersti Börjars, Manchester
Kaja Borthen, Trondheim
Kjersti Maria Rongen Breivega, Bergen
Leiv Egil Breivik, Bergen
Joan Bresnan, Stanford
Endre Brunstad, Bergen
Edit Bugge, Bergen
Trude Bukve, Bergen
Tove Bull, Tromsø
Tore Burheim, Bergen
Miriam Butt, Konstanz
Aoife Cahill, Princeton
Margrete Dyvik Cardona, Bergen
Mauricio Cardona, Laksevåg
Özlem Çetinoğlu, Stuttgart
Ana B. Chiquito, Bergen
Kirsti Koch Christensen, Oslo
Robin Cooper, Gothenburg
Trine Dahl, Bergen
Eystein Dahl, Tromsø
Östen Dahl, Stockholm
Mary Dalrymple, Oxford
Valeria de Paiva, Cupertino
Koenraad De Smedt, Bergen
Cheikh Bamba Dione, Bergen
Tore Dyvik, Trondheim
Dårdi Øye Dyvik, Trondheim
Einar Dyvik, Bergen
Hanne Dyvik, Stavanger
Signe Oksefjell Ebeling, Oslo
Hanne Eckhoff, Oxford
Øystein Eek, Bærum
Kristin Melum Eide, Trondheim
Elisabet Engdahl, Göteborg
Hans-Olav Enger, Oslo
Hanne Erdman Thomsen, København

Kristin Espedal, Bergen
Jan Terje Faarlund, Oslo
Cathrine Fabricius-Hansen, Oslo
Jens Erik Fenstad, Oslo
Ruth Vatvedt Fjeld, Oslo
Dan Flickinger, Stanford
Kjersti Fløttum, Bergen
Martin Forst, Heidelberg
Anette Frank, Heidelberg
Kari Fraurud, Edsbro
Siri Fredrikson, Bergen
Thorstein Fretheim, Trondheim
Jan Olav Fretland, Lærdal
Jan Olav Gatland, Bergen
Anje Müller Gjesdal, Bergen
Anne Golden, Oslo
Atle Grønn, Oslo
Oddrun Grønvik, Oslo
Heming Helland Gujord, Fana
Ann-Kristin Helland Gujord, Fana
Tor Guttu, Oslo
Margareth Hagen, Bergen
Kristin Hagen, Oslo
Jan Ragnar Hagland, Trondheim
Odile Halmøy, Bergen
Madeleine Halmøy, Fimreite
Sandra L. Halverson, Bergen
Hans Marius Hansteen, Bergen
Lidun Hareide, Flø
Vidar Haslum, Kristiansand
Hilde Hasselgård, Oslo
Odd Einar Haugen, Bergen
Petter Haugereid, Bergen
Lilian Haugereid, Bergen
Kari Haugland, Bergen
Åsta Haukås, Bergen
Dagmar Haumann, Bergen
Annette Hautli-Janisz, Konstanz
Eldar Heide, Bergen
Mikael Heimann, Göteborg
Katarina Heimann Mühlenbock, Gothenburg
Stig Jarle Helset, Volda
Kjetil Berg Henjum, Bergen
Petter Henriksen, Oslo
Frøydis Hertzberg, Oslo
Torill Hestetræet, Bergen
Johs. Hjellbrekke, Bergen
Knut Hofland, Bergen
Jan Kristian Hognestad, Stavanger
Helge Vidar Holm, Bergen
Marit Hovdenak, Haslum
Olaf Husby, Trondheim
Nancy Ide, Poughkeepsie
Þorsteinn G. Indriðason, Bergen
Benedicte Irgens, Bergen
Ernst Håkon Jahr, Kristiansand
Bård Uri Jensen, Hamar
Ole-Jørgen Johannessen, Bergen
Christer Johansson, Bergen
Åse Johnsen, Bergen
Marit Julien, Lund
Daniel Jung, Bergen
Annette Myre Jørgensen, Østfold
Hans Kamp, Stuttgart
Ronald M. Kaplan, Palo Alto
Lauri Karttunen, Emerald Hills
Martin Kay, Menlo Park
Paul Kerswill, York
Anna Kibort, Oxford
Tracy King, Mountain View
Torodd Kinn, Os
Marit Helene Kløve, Bergen
James Knirk, Oslo
Eli Kristine Knudsen, Bergen
Nazuki Kobayashi, Bergen
Eivind Kolflaath, Bergen
Randi Korneliussen, Bergen
Kimmo Koskenniemi, Helsinki
Atle Kristiansen, Hagavik

Gjert Kristoffersen, Bergen
Jonas Kuhn, Stuttgart
Lars Anders Kulbrandstad, Hamar
Rune Kyrkjebø, Fyllingsdalen
Tore Langholm, Bergen
Natascia Leonardi, Macerata
Ragnhild Lie Anderson, Bergen
Krister Lindén, Helsinki
Arild Linneberg, Bergen
Terje Lohndal, Trondheim
Gyri Smørdal Losnegaard, Straumsgrend
Ingunn Lunde, Bergen
Bente Luneng, Bergen
Gunn Inger Lyse, Bergen
Helge Lødrup, Oslo
Jan Tore Lønning, Oslo
Bente Maegaard, Copenhagen
Joan Maling, Arlington
John Maxwell, Palo Alto
Paul Meurer, Bergen
Johan Myking, Bergen
Klaus Johan Myrvoll, Oslo
Brit Mæhlum, Trondheim
Marit Aamodt Nielsen, Kristiansand
Randi Alice Nilsen, Trondheim
Ingvild Nistov, Bergen
Joakim Nivre, Uppsala
Torbjørn Nordgård, Trondheim
Anders Nøklestad, Oslo
Stephan Oepen, Oslo
Lise Opdahl, Bergen
Christian-Emil Ore, Oslo
Petya Osenova, Sofia
Carla Parra Escartín, Dublin
Marco Passarotti, Milan
Agnieszka Patejuk, Warsaw
Martin Paulsen, Bergen
Håvard Peersen, Bergen
Alois Pichler, Bergen
Richard H. Pierce, Bergen
Adam Przepiórkowski, Warsaw
Silje Ragnhildstveit, Bergen
Margunn Rauset, Bergen
Øystein Reigem, Florvåg
Jill Walker Rettberg, Bergen
Tomas Riad, Stockholm
Curt Rice, Oslo
Jan Roald, Bergen
Eiríkur Rögnvaldsson, Reykjavík
Victoria Rosén, Bergen
Rune Røsstad, Søgne
Unn Røyneland, Oslo
Louisa Sadler, Wivenhoe
Anna Sågvall Hein, Uppsala
Christine Meklenborg Salvesen, Oslo
Andrew Salway, Bergen
Helge Sandøy, Bergen
Jens Eike Schnall, Bergen
Jørgen Magnus Sejersted, Bergen
Peter Sells, York
Gisle Selnes, Bergen
Helle Frisak Sem, Oslo
Hanne Gram Simonsen, Oslo
Halvard Sivertsen, Bergen
Klara Sjo, Bergen
Per Erik Solberg, Oslo
Dansk Sprognævn, Frederiksberg
Kjetil Strand, Oslo
Sebastian Sulger, Konstanz
Andreas Sveen, Oslo
Torbjørn Svendsen, Trondheim
Arne S. Svindland, Bergen
Željka Švrljuga, Bergen
Toril Swan, Bergen
Kjell Johan Sæbø, Oslo
Lars Sætre, Bergen
Sindre Sørensen, Bergen
Knut Tarald Taraldsen, Tromsø
Kari Tenfjord, Bergen
Rolf Theil, Jevnaker

Martha Thunes, Bergen
Jörg Tiedemann, Helsinki
Ingebjørg Tonne, Oslo
Ludmila Torlakova, Kleppestø
Trond Trosterud, Tromsø
Dag Trygve Truslew Haug, Oslo
Ivar Utne, Bergen
Karit Elise Valen, Bergen
Øystein A. Vangsnes, Tromsø
Erik Velldal, Oslo
Åke Viberg, Stockholm
Lars Vikør, Oslo
Arnfinn Muruvik Vonen, Oslo
Atro Voutilainen, Helsinki
Boye Wangensteen, Kurland
Jürgen Wedekind, Copenhagen
Eirik Welo, Oslo
Marit Westergaard, Tromsø
Åse Wetås, Drammen
Anssi Yli-Jyrä, Helsinki
Gisle Ytrestøl, Oslo
Annie Zaenen, Emerald Hills
Lilja Øvrelid, Oslo
Tor A. Åfarli, Trondheim

Contributors

Kristin Bech

- *University of Oslo*

Kristin Bech is associate professor of English language. She specializes in the development of syntactic structures in the history of English, with a focus on Old and Middle English, but she is also interested in cross-linguistic comparisons. She has previously run a project on the relation between syntax and information structure in old Germanic and Romance languages, and has recently started a new project on the structure of noun phrases in old Germanic languages.

Kaja Borthen

- *Norwegian University of Science and Technology*

Kaja Borthen is professor in linguistics at the section for Scandinavian Languages of Department of Language and Literature. She has worked on syntactic, semantic and pragmatic aspects of nominal phrases, using Head-driven Phrase Structure Grammar as her main syntactic framework, and on topics related to automatic anaphora resolution. Her most recent research has focused on non-truth conditional semantics and pragmatics, including topics such as implicature, the semantics and pragmatics of referring expressions, and the semantics and pragmatics of pragmatic particles.

Robin Cooper

- *University of Gothenburg*

Robin Cooper is senior professor at the University of Gothenburg, where he was previously professor of computational linguistics. He is currently conducting research within the Centre for Linguistic Theory and Studies in Probability (CLASP) at the Department of Philosophy, Linguistics and Theory of Science. His main research interests are semantics (both theoretical and computational), dialogue semantics and computational dialogue systems. Currently he is working on a type theoretical approach to language and cognition.

Mary Dalrymple
■ *University of Oxford*
Mary Dalrymple is professor of syntax at the Faculty of Linguistics. Her research centers on syntax, the syntax-semantics interface, and semantics, particularly within the framework of Lexical Functional Grammar. She is interested in the syntactic properties of human languages and how they can guide the process of assembling meanings of words and phrases into meanings of larger phrases and sentences. She is also interested in language description and documentation, and in Austronesian and Papuan languages.

Cheikh M. Bamba Dione
■ *University of Bergen*
Cheikh M. Bamba Dione is lecturer in linguistics and computational linguistics at the Department of Linguistic, Literary and Aesthetic Studies. His interests include theoretical and computational linguistics as well as psycholinguistics. He has done extensive work on Wolof morphology and grammar. His most recent research focused on the implementation of computational models to develop digital language resources. He is also conducting studies on the use of statistical and machine learning approaches to develop language technology applications. In the area of psycholinguistics, his research interests are language comprehension and production, interaction of language and thought, and mechanisms that influence human language representation and processing.

Elisabet Engdahl
■ *University of Gothenburg*
Elisabet Engdahl is professor emerita at the Department of Swedish. She is interested in Scandinavian linguistics, in particular variation in word order and information structure. She has been involved in the Scandinavian Dialect Syntax project and in the development of the Nordic Dialect Corpus and the Nordic Syntax Database, two internet resources maintained at the Text Laboratory at the University of Oslo. With this article she resumes an old collaboration with Annie Zaenen and Joan Maling, dating back to a joint paper published in 1981.

Dag Trygve Truslew Haug
■ *University of Oslo*
Dag Haug is a professor of Greek and Latin at the Department of Philosophy, Classics, History of Art and Ideas. His work focuses on formal syntax (using Lexical Functional Grammar) and semantics, as well as historical linguistics. He works both on the classical languages and on issues of more general theoretical interest, in particular anaphora, control and binding. He leads the development of the PROIEL treebank of ancient Indo-

European languages, which aims to offer scholars who work on these languages with a solid empirical basis for their research.

Petter Haugereid

- *Western Norway University of Applied Sciences*

Petter Haugereid is associate professor in the Norwegian Department, teaching Norwegian since 2014. He has had postdoc positions at NTU, Singapore from 2010 to 2012, Haifa University from 2012 to 2013 and the University of Bergen from 2013 to 2014. He has worked on machine translation, computational grammars and syntactic annotation of Norwegian corpora. His thesis at NTNU, Trondheim, from 2009 was on phrasal subconstructions and presented a constructionalist grammar design, exemplified with Norwegian and English. His publications are mainly in the field of formal grammar, machine translation, and corpora.

Christer Johansson

- *University of Bergen*

Christer Johansson is professor of computational linguistics. Previously he was a post-doc researcher at the Institute of Advanced Industrial Science and Technology in Tsukuba, Japan, and a member of Laurie Ann Stowe's research group at Rijksuniversiteit Groningen. His doctoral dissertation from Lund University in 1997 modeled learnability of language. His research interests include cognitive science, experimental psycholinguistics and statistical models of language. His teaching aims to combine research and teaching, which has resulted in several MA-theses and conference contributions on themes related to crosslinguistic priming, code switching, reference as well as theory testing.

Ronald M. Kaplan

- *Amazon.com and Stanford University*

Ronald M. Kaplan is currently Chief Scientist for Search Technologies at Amazon.com and adjunct professor of linguistics at Stanford University. For many years he directed the Natural Language Theory and Technology research group at the Xerox Palo Alto Research Center. He created the modular architecture of Lexical Functional Grammar and introduced many of its formal devices for linguistic description. His research centers on the mathematical and computational properties of the LFG formalism and its ability to support well-motivated accounts for a wide range of linguistic phenomena.

Torodd Kinn

- *University of Bergen*

Torodd Kinn is professor of Scandinavian linguistics. His research has three foci. The first is in Scandinavian morphology and syntax, with studies centered on pseudopar-

titives (binominal constructions), grammaticalization, and pseudocoordination (complex predication). The second focus is in text linguistics, with studies of scientific writing (especially prononun use) in Norwegian, English, and French research articles in the fields of economics, medicine, and linguistics. The third focus is in anthroponymy, with the development of innovative methods in historical Norwegian given-name geography.

Helge Lødrup

■ *University of Oslo*

Helge Lødrup is professor of general linguistics in the Department of Linguistics and Scandinavian Studies. His field of research is the grammar of Norwegian and related languages within generative grammar, especially Lexical Functional Grammar. He has published on various topics, mainly within Norwegian syntax, such as passive and impersonal sentences, reflexives, body part nouns and kinship nouns, external possessors and other possessive expressions, surface anaphora, clausal complementation, complex predicates, and pseudocoordination.

Joan Maling

■ *Brandeis University and National Science Foundation*

Joan Maling is professor emerita of linguistics in the Linguistics Program, where she taught from 1972 until 2003. Since June 2003 she has been director of the Linguistics Program at the National Science Foundation, where she was instrumental in starting the Documenting Endangered Languages Program. She has published on many aspects of the syntax of Modern Icelandic, especially case, word order, passive, preposition-stranding and long distance reflexives, and on case alternations in Finnish, Korean and German as well as Icelandic. In 2009, she was awarded an honorary doctorate by the University of Iceland for her contributions to Icelandic linguistics.

Paul Meurer

■ *Uni Research Computing and University of Bergen*

Paul Meurer is a researcher at Uni Research Computing and at the University of Bergen, Department of Linguistic, Literary and Aesthetic Studies. His research interests lie in the fields of theoretical and computational linguistics. Within theoretical linguistics, he has focused on morphology and syntax, and language typology. In computational linguistics, he has contributed to the research and development of language resources and tools in diverse fields, such as morphological and syntactic parsing for Norwegian, Georgian and Abkhaz, treebanking, vizualization, corpus management and search, terminology, and metadata curation. He received the Steven Krauwer Award for CLARIN Achievements in 2017.

Adam Przepiórkowski
■ *Polish Academy of Sciences and University of Warsaw*
Adam Przepiórkowski holds the professor position both at the Institute of Computer Science of the Polish Academy of Sciences and at the Institute of Philosophy of the University of Warsaw, where he teaches linguistics at the Cognitive Science programme. He headed the National Corpus of Polish project, he worked within Head-driven Phrase Structure Grammar on case, negation, etc., and he currently works within Lexical Functional Grammar on the argument–adjunct (non)distinction, on coordination, distributivity, case assignment and other issues at the syntax-semantics interface.

Victoria Rosén
■ *University of Bergen*
Victoria Rosén is associate professor of linguistics at the Department of Linguistic, Literary and Aesthetic Studies. Her main research interest is syntax in the Lexical Functional Grammar framework, and her dissertation concerned an LFG analysis of topics and empty pronouns in Vietnamese. In later work she has focused on the grammar of Norwegian in language technology projects dealing with automatic proofreading, machine translation, computational grammar and treebanking. She led the INESS treebanking project, which established an infrastructure for treebanking and created a large LFG treebank for Norwegian. Recently she has been working with multiword expressions and how they are represented in treebanks.

Martha Thunes
■ *Western Norway University of Applied Sciences*
Martha Thunes is associate professor of English. She is a general linguist with a background from the University of Bergen. Her research has been centred around English-Norwegian contrastive language analysis, combined with theoretical and computational linguistics, translation theory, corpus linguistics, text typology, and studies of language for special purposes. She has worked with treebanking for Norwegian and contributed to NorGramBank. She has published works within natural language processing, general linguistics and contrastive language studies, and her main scientific interests are contrastive linguistics, grammar and the lexicon.

Annie Zaenen
■ *Stanford University*
Annie Zaenen is adjunct professor in linguistics and also a researcher at the Center for the Study of Language and Information. She is retired from Xerox PARC. She is interested in syntax, lexical semantics, formal issues in grammar and the linguistic side of knowledge representation. She has a theoretical affiliation with Lexical Func-

tional Grammar but is mainly interested in getting the right generalizations about the interplay of structural constraints, lexical roles, discourse and information structure.

Preface

Helge Dyvik will be 70 on December 23, 2017. We are proud to present this volume of linguistic studies by colleagues from near and far who have welcomed the opportunity to honor Helge and his career. The title is a take on the song 'I Am the Very Model of a Modern Major-General' from the Gilbert & Sullivan comic opera *The Pirates of Penzance*. Although the song, which Helge knows and enjoys, is a parody packed with hyperbole, our admiration of Helge's academic achievements can hardly be overstated, for reasons which will soon become apparent.

Helge grew up in Bodø in northern Norway. Already when he was quite young, he showed a special interest in languages. He sometimes wrote his homework assignments in verse, and he was in fact a bit of a problem for his language teachers, because he knew more than they did. He had a special fascination for Professor Henry Higgins in George Bernard Shaw's *Pygmalion*, and he read Otto Jespersen's *The Philosophy of Grammar* while at school — early signs that he was headed towards a life in linguistics. Once he was done with school, he could hardly wait to begin his university studies, so he spent a lot of his free time during his military service studying Latin, much to the amazement of his fellow soldiers.

Once finished with school and the military, Helge left provincial Bodø and traveled south. He felt he had come to the big city when he arrived at the University of Bergen to commence his studies, and his first subject was phonetics. His undergraduate degree also included English and Scandinavian language and literature. He received a bachelor's degree in 1972 and continued studying Scandinavian languages for his graduate degree in 1976. His studies also included a spell at the University of Durham in England, where he studied Old English language and literature and enjoyed wearing a gown to the formal dinners at the dining hall. Back in Bergen he studied Vietnamese and Cantonese.

From 1974 to 1981 Helge was employed at the University of Bergen as a research assistant and later lecturer in Old Norse, and during this period he participated in a project on the grammar of the language. It was his responsibility to treat the syntax of Old Norse, and his intention was to write his doctoral dissertation on this topic. The first sentence of his dissertation is 'Er lingvistikken en empirisk vitenskap?' (Is linguistics an empirical science?), and the reader can perhaps guess where it goes from there. Helge wanted a modern linguistic framework for describing the syntax of the

old language, but found that the transformational generative grammar of the time did not measure up. Rather than being about Old Norse syntax, the dissertation ended up proposing a new model for grammatical description, one which had much in common with Lexical Functional Grammar, which was being developed at the same time. If Helge's dissertation had been written in English rather than Norwegian, we feel certain that it would have been influential on the international stage.

In 1983 Helge became professor of general linguistics. During his career he has had an unusually broad range of interests. He has done research on Old Norse and Old English phonology (umlaut and breaking), Old Norse syntax (passive and the development of articles), and runology (interpretation of runic inscriptions). He has also studied Vietnamese syntax (classifiers and topic constructions). Throughout his career he has been engaged in foundational issues in linguistics.

In the 1980s Helge became interested in computational linguistics. In the late 1980s and early 1990s he developed PONS, an experimental system for machine translation. The system took advantage of structural similarities between the source and target languages to take shortcuts during the translation process, thereby achieving a compromise between linguistic sophistication and efficiency.

From 2001 to 2004 he led the project From Parallel Corpus to Wordnet in which the *Semantic Mirrors* method for deriving semantic information from translations was developed. Based on the assumptions that semantically closely related words should have strongly overlapping sets of translations, and that words with wide meanings should have a higher number of translations than words with narrow meanings, the method formulated definitions for semantic concepts such as 'synonymy', 'hyponymy', 'ambiguity' and 'semantic field' in translational terms.

In 1999 Helge initiated Norwegian participation in the international Parallel Grammar project, and he led the development of NorGram, the Norwegian ParGram grammar based on Lexical Functional Grammar. After the initial NorGram project, the grammar has been used in many other projects. From 2003 to 2007 Helge led the Bergen group participating in the LOGON machine translation project, in which translation was done not only from one language to another (from Norwegian to English), but also from one syntactic framework to another (from LFG to HPSG). For this project Helge added a Minimal Recursion Semantics projection to NorGram to enable semantic transfer-based translation. NorGram has also been applied in the projects TREPIL (the Norwegian Treebank Pilot Project) and XPAR (Language Diversity and Parallel Grammars); the latter formulated formal principles for aligning monolingual treebanks at phrase and word levels based on translational correspondences at predicate-argument level.

Helge played a leading role in the INESS project (Infrastructure for the Exploration of Syntax and Semantics), which ran from 2010 to 2017. This project created NorGramBank, a large treebank for Norwegian Bokmål and Nynorsk, by parsing a corpus auto-

matically with NorGram. Helge was responsible for the further development of NorGram throughout the project, and his insightful analyses of practically every syntactic construction in the language has resulted in a treebank with very detailed syntactic annotation. The combination of Helge's thorough understanding of all aspects of Norwegian grammar and his extraordinary talent for implementation has resulted in a computational grammar of great sophistication.

Helge has had many collaborations with colleagues outside of Bergen. In the academic year 1996/1997 he was affiliated with the research group *Contrastive Analysis and Translation Studies Linked to Text Corpora* at the Centre for Advanced Study in Oslo. He spent several sabbaticals at the Palo Alto Research Center in California and developed good relations with its natural language research group. He has recently done research on Norwegian language subnorms and has become involved in the BRO dictionary project. He is currently affiliated with the research group *SynSem: From Form to Meaning — Integrating Linguistics and Computing* at the Centre for Advanced Studies in Oslo.

Helge has had a number of responsibilities outside of his duties at the University of Bergen. One of his most important appointments was as the chair of the Language Council of Norway's Committee for Language Standardization and Language Observation, a position in which he served from 2007 to 2014. The Language Council is the state advisory body for the Norwegian language in both its written forms, Bokmål and Nynorsk. The committee worked on a variety of issues concerning language standardization, revision of the orthography of Norwegian, the management and future of monolingual lexicography for Norwegian, nondiscriminatory language in the media, etc. Helge's approach in dealing with all these issues was that all decisions should be based on explicit reasoning, and that every practical decision should be anchored in linguistic theory in a principled way.

Unlike some of his younger colleagues, Helge is active in social media, especially Facebook. *Språkspalta* (The Language Column) is a Facebook group which counts some 34,000 members, many of whom have very strong opinions about the Norwegian language – opinions that are not necessarily based on facts or research. Helge is a veritable beacon of enlightenment in this environment that most linguists shun like the plague. He is untiring in his attempts to educate this group about language and linguistics, and there is hardly any question or statement he is not able to offer insightful comments on. His contributions about etymology and language history are especially valued. Helge's posts in *Språkspalta* are concise and highly informative, often with a touch of humor, though he does sometimes get exasperated, especially when he has to repeat — again and again — that languages change naturally, or that Bokmål is not Danish, or that spoken language and written language are two different things. Faithful members of *Språkspalta* describe him as a hero and a guru, eminent and elegant, practical and pedagogical, a knight in shining armor battling ignorance and prejudice.

Helge is well known for his sense of humor. When in working with NorGram we sometimes came upon a feature of the Norwegian language that was practically impossible to find a good solution for, Helge would often joke that he would probably be better off just changing the language, something he could easily do because of his influence with the Language Council. Helge has fun with language whenever he has a chance. Puns, plays on words, rhymes, every way of playing around with languages comes naturally to him all the time. His poems, mostly produced for and performed at special academic occasions, have become famous, and not just at our department. When the University of Bergen had its fiftieth anniversary as a university in 1996, Helge was commissioned to write and perform a prologue for the celebration. For this occasion he wrote nearly 2000 words in 370 lines of verse. Helge dismisses these writings as *rimerier* (rhymings), but we who know him well, know better; if he hadn't chosen to be a linguist, he could equally well have been a poet.

It is with a certain melancholy that we celebrate Helge's career with this festschrift. On the one hand, we are happy to congratulate him on (soon) reaching the ripe old age of 70. On the other hand, it is hard to imagine our work environment without him, and we will certainly miss his important contributions to the department's teaching, supervision and administrative work. But luckily for us, he will not be disappearing. He will be transformed from professor to professor emeritus, but since he's never really believed in transformations, we don't expect this to have much of an effect on his work ethic. We look forward to having him as a colleague and friend for many years to come.

We would like to thank our distinguished international panel of reviewers for their work with the articles in this volume. We are also deeply grateful to Kristin Bech for her expert help with proofreading. Finally we thank Kristin Bech, Oddrun Grønvik, Halvard Sivertsen, Martha Thunes and others for help with the preface.

Bergen, November 23, 2017 Victoria Rosén and Koenraad De Smedt

Old English and Old Norwegian noun phrases with two attributive adjectives

Kristin Bech

Abstract. The topic of this paper is Old English and Old Norwegian noun phrases containing two attributive adjectives. An overview of the frequency of various word order constellations will be given, before we zoom in on one of them, namely the construction Adjective – Adjective – Noun, i.e. noun phrases in which two prenominal adjectives occur next to each other without a coordinating conjunction. Old English and Old Norwegian will be compared with respect to which adjectives occur in this position. The paper also includes an intermezzo, during which we investigate what happens to adjective position when a text is translated from present-day English into Old English.

1 Introduction

Old English and Old Norwegian are closely related early Germanic languages. Although a few centuries separate them with respect to the written record – Old Norwegian was not written down (in Latin script) until the thirteenth century, whereas the Old English written tradition started in the ninth century – they can nevertheless be said to represent approximately the same early Germanic stage. In fact, according to Lass (2000), who compared ten features in related Germanic languages with the purpose of placing them on a scale from less archaic to more archaic, Old Norse[1] is slightly more archaic than Old English.

With the notable exception of Old (and Middle) English (Mitchell 1985 for basic taxonomy; Fischer 2000, 2001, 2006, 2012; Fischer and van der Wurff 2006; Haumann 2003, 2010; Pysz 2007, 2009), noun phrase structure in early Germanic languages is an under-researched area, especially in a cross-linguistic perspective.[2] Claims about

1 I.e. Old Norwegian and Old Icelandic. Lass mentions Old Icelandic, not Old Norse, but we can assume that Old Norwegian belongs there as well. Old Icelandic and Old Norse are sometimes – erroneously – used as synonyms.

2 The situation can be expected to be remedied in the next few years. The project 'Constraints on syntactic variation: noun phrases in early Germanic languages', funded by the Research Council of Norway, runs from September 2017 to August 2020.

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1326.

adjective position in Old Norse, stemming from Nygaard (1906) and Ringdal (1918), have been repeated in the century since (e.g. Valfells and Cathey 1981; Haugen 1995; Faarlund 2004), but as received wisdom rather than as possible research topics. As regards Old English, Fischer and Haumann have taken different positions, with Fischer arguing that there is a connection between adjective position on the one hand, and definiteness, declension and linear iconicity on the other (see e.g. Fischer 2000, p. 170; 2012, p. 252), whereas Haumann's stance is that adjective position 'follows exclusively from interpretive and functional differences' (2010, p. 54). Pysz's (2009) main concern is to account for the observed variation within a Chomskyan generative framework. Their suggestions will be discussed in future papers. The purpose of this paper is to give an empirical overview of adjective position in Old English and Old Norwegian noun phrases containing two attributive adjectives.[3]

Noun phrases containing one adjective are generally common in the old languages,[4] but unlike the present-day languages, the old languages did not favour clusters of adjectives within the noun phrase; noun phrases with two adjectives are not particularly frequent, as we shall see, and more than two adjectives are rare within a noun phrase (see also Pysz 2009, p. 29). In present-day English and Norwegian, adjectives can easily be stacked, but it happens in a certain order, depending on the semantic properties of the adjective (see e.g. Quirk et al. 1985, p. 1337ff; Faarlund et al. 1997, p. 407–410). For example, non-gradable intensifiers occur before gradable adjectives, which occur before participles and colour adjectives, which occur before nationality adjectives. Hence, example (1) is perfectly fine, whereas (2) would be odd. Furthermore, postnominal adjectives are rare in the present-day languages; notable exceptions are set phrases, often loans, as in (3), phrases in which the head is an indefinite pronoun, such as (4), and phrases with modified adjectives, such as (5)[5] (Quirk et al. 1985, p. 1293 f.).

(1) a certain intelligent retired Norwegian professor

(2) a certain Norwegian retired intelligent professor

(3) The professor emeritus mostly lived off spaghetti bolognese in his retirement.

(4) somebody nice

(5) a mistake typical of absent-minded professors

3 One of Fischer's (2000, 2001, 2006, 2012) main points is that postnominal adjectives are 'functionally predicative' even when they are not in a predicative construction with a copula. I will not consider that proposal here; in this study I have regarded as attributive all adjectives that are annotated as modifying a head noun.

4 A simple query for noun phrases containing one adjective gave 42,291 hits in the Old English corpus and 5,048 in the Old Norwegian corpus – compare with the numbers in Table 1.

5 Norwegian does not have an equivalent of *somebody nice*, and as regards example (5), the construction is marginally possible in Norwegian: *en feil typisk for distré professorer*, but Norwegian would prefer to use a relative clause instead: *en feil som er typisk for distré professorer*.

Old English and Old Norwegian allowed postnominal adjectives to a much greater extent, including descriptive adjectives. In other words, the present-day languages differ quite considerably from the early languages with respect to adjectives in noun phrases, and this paper describes and discusses some of the old patterns.

2 Method

The data is taken from two corpora, the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (YCOE, Taylor et al. 2003), and the *Menotec* corpus of Old Norwegian, hosted by the INESS infrastructure (Rosén et al. 2012). YCOE contains c. 1.5 million words in 100 texts of different genres from both early and late Old English (c. 800–1100). Menotec is a much smaller corpus, consisting of c. 214,000 words in four thirteenth century texts of different genres. These are *The Old Norwegian homily book, The legendary saga of St. Olaf, Magnus Lagabøte's national law code,* and *Strengleikar* 'stringed instruments', a chivalric saga adapted from Old French. Hence the question arises as to whether data from these corpora can be compared at all. I will argue that they can, for the reason that noun phrases containing attributive adjectives are generally frequent. Consequently, even a small corpus can be expected to yield general patterns. It will also become clear that Old English and Old Norwegian are in many respects quite similar as concerns noun phrase structure, which is what we would expect in two closely related languages. In other words, in terms of the data distribution, the situation is reassuring with respect to comparability. However, as in all corpus work, we nevertheless proceed with caution, keeping an eye out for possible genre differences, especially since the range of genres is much wider in YCOE than in Menotec.

I searched for all noun phrases containing two attributive adjectives.[6] This was done by means of a number of queries, which also differ depending on the corpus, since YCOE annotates phrase structure in the Penn Treebank format, whereas Menotec annotates dependency structure. It was therefore a challenge to write the queries in such a way that they would yield comparable patterns for each category in the two languages. Consequently, I started out with relatively general queries, studied the results, and then narrowed the queries gradually. The different ordering possibilities are presented in Section 3. Note that this study is intended as an overview for the purpose of a general comparison of Old English and Old Norwegian, hence there are some distinctions that have not been made. For example, I have not distinguished between strong and weak adjectives, or between noun phrases with or without determiners. Some of the complexities are commented on in connection with each pattern, as the examples chosen for illustration are usually the most 'bare' examples.

6 Note that adjectival participles have not been included.

3 General overview of ordering possibilities

Table 1 shows the results for Old English and Old Norwegian. The two languages are similar in the sense that all the patterns are possible, and they are also fairly similar with respect to the overall distribution. The most common patterns are A-A-N and A-N-*and*-A in both languages, accounting for 64.5% of the instances in Old English and 56.1% of the instances in Old Norwegian. Old Norwegian seems to favour explicit postnominal coordination more than Old English does, while Old English to a greater extent flanks the adjectives (A-N-A) (but see comments below).

	Old English		Old Norwegian	
	#	%	#	%
A-A-N	296	29.5	51	26.0
A-N-A	108	10.8	9	4.6
N-A-A	2	0.2	11	5.6
A-*and*-A-N	214	21.3	43	21.9
A-N-*and*-A	351	35.0	59	30.1
N-A-*and*-A	33	3.3	23	11.7
Total	1004	100.1	196	99.9

Table 1: Distribution of adjectives in noun phrases with two adjectives in Old English and Old Norwegian

Examples of the various constellations are given in (6)–(11), with Old English in the *a* examples and Old Norwegian in the *b* examples.[7]

3.1 Patterns without conjunction

A-A-N

(6) *a.* ***ealdum*** ***leasum*** *spellum* (coboeth: 35.98.25.1907)
old false tales
'old false tales'

b. *einn* ***ricr*** ***lenndr*** *maðr* (strleik: 2191)
a rich landholding man
'a rich landholding man'

The A-A-N pattern, which is the focus of the present paper, will be considered in greater detail in Section 5.

7 For readability, Old Norwegian ‹v› and ‹ſ› have been normalized to ‹u› and ‹s›. For example, *miclu* in example (7) reads *miclv* in the corpus. The source for each example is provided with the codes used in the corpora.

A-N-A

(7) *a.* ***medmicle*** *nose* ***þynne*** (cobede: 13.144.14.1391)
moderate nose thin
'moderate thin nose'

b. ***miclu*** *lofte* ***steinþildu*** (strleik: 2191)
large loft stone-tiled
'large stone-tiled loft'

Whereas the modern languages would place the adjectives prenominally, the old languages could easily postpone one of them, either with (A-N-*and*-A) or without (A-N-A) a conjunction. The A-N-A pattern is more frequent in Old English than in Old Norwegian, but here it should be noted that almost half of the Old English occurrences are from two medical handbooks, which use certain constructions to describe the ingredients needed in the various recipes for treatment. If these texts are removed from the corpus, A-N-A is still more frequent in Old English than in Old Norwegian, but the difference is much less pronounced. We will shortly consider the A-A-N pattern in some detail, and it will then become clear that although both languages allowed two prenominal adjectives, there were restrictions on which adjective types could occur together in prenominal position. In other words, there was a reason for the postponement of some adjectives.

N-A-A

(8) *a.* *wingeardes* *twigu* ***ufeweard*** ***merwe*** (colaece: 12.1.5.2273)
vine twig upper tender
'a vine twig of which the upper part is tender'

b. *systir* ***samfæðra*** ***skilgeten*** (mll: 928)
sister same-father.ADJ trueborn
'sister, trueborn of the same father'

Noun phrases which contain two postnominal adjectives that are not linked by a coordinating conjunction are rare, especially in Old English, and the few examples that exist are formulaic expressions. All except one of the Old Norwegian examples are from *Magnus Lagabøte's national law code.*

3.2 Patterns with conjunction

A-*and*-A-N

(9) *a.* ***torhtum*** *&* ***swutolum*** *wordum* (cogregdC: 36.175.2.2130)
clear and plain words
'clear and plain words'

b. ***margar*** *oc* ***rikar*** *giaver* (strleik: 2017)
large and rich gifts
'large and rich gifts'

The A-*and*-A-N category also includes noun phrases in which a determiner precedes the first adjective or both adjectives. This leads to some issues concerning the interpretation of the data, which will be discussed below.

A-N-*and*-A

(10) *a.* ***grimlicre*** *stefne ond* ***ladlicre*** (comart3: Au25,A.9.1538)
fierce voice and unpleasant
'fierce and unpleasant voice'

b. ***agiætleg*** *takn oc* ***fagrleg*** (olavssaga: 2644)
excellent sign and beauteous
'excellent and beauteous sign'

As regards this pattern, Old English and Old Norwegian are quite similar in terms of distribution, and this is perhaps a construction that we typically associate with Old English and Old Norse. Here as well, a determiner may precede one or both adjectives, cf. discussion below.

N-A-*and*-A

(11) *a.* *ða maðmfatu* ***gyldene*** *and* ***sylfrene*** (cocathom2: 33:252.100.5621)
the costly vessels golden and silvern
'the costly vessels of gold and silver'

b. *maðr* ***spakr*** *oc* ***fastnæmr*** (olavssaga: 191)
man wise and faithful
'a wise and faithful man'

In this construction, a noun is followed by two coordinated adjectives. For the purposes of this paper I have only considered noun phrases with two adjectives. Although they are infrequent, constructions with more than two postnominal adjectives exist, especially in Old Norwegian, it seems. This could be due to the genres included. For example, one of the texts in the Old Norwegian corpus is the chivalric saga *Strengleikar*. Hence, we get descriptions like the ones in (12) and (13), which serve to flavour the story.

(12) *einn riddare* ***curteis*** *oc* ***vaskr*** *oc* ***vapndiarfr*** (strleik: 2433)
a knight courtly and able and weapon-brave
'a courtly and able knight, brave with weapons'

(13) *grimm kona oc* ***drambsom illmalog*** *oc* ***ovundsiuk*** (strleik: 424)
grim woman and arrogant ill-spoken and jealous
'a grim, arrogant, ill-spoken and jealous woman'

It was mentioned above that as regards the patterns with conjunction, a determiner may precede one or both adjectives. If the second adjective is preceded by a determiner, the reference is more likely to be what Fischer (2012, p. 266–267) terms 'sloppy'; i.e. the adjectives do not refer to the same entity, especially in the A-N-*and*-A pattern. Fischer only considers Old English, but we can assume that the same is the case in Old Norse, although more work needs to be done here. An Old English example is given in (14), where it is clear that it is not the same citizens that are good and evil.[8] This has implications for the formal analysis of the phrases, which future work will have to take into account.

(14) *þa* ***godan*** *ceastergewaran and ða* ***yfelan*** (cocathom2: 4:38.262.854)
the good citizens and the evil
'the good and the evil citizens'

The reference can also be 'strict' in phrases with two determiners, especially with singular head nouns. An example is (15), where the two adjectives refer to the same king. The proportion of strict identity in such phrases is lower than in phrases without a second determiner, but that does not mean that this kind of reference is rare: 46.1% of Fischer's cases in postnominal *and*-constructions had strict identity (2012, p. 267).[9]

(15) *se* ***strongesta*** *cyning & se* ***gylpgeornesta****, Æðelfrið haten*
the mightiest king and the proudest, Æthelfrith called
'the mightiest and proudest king, called Æthelfrith' (cobede: 1:18.92.3.838)

It also happens, though not frequently, that the reference is sloppy in constructions without determiners, cf. example (16) from Old Norwegian, where the reference is obviously to different men. For Old English, Fischer (2012, p. 267) reports a proportion of 4.3% sloppy identity in these constructions.[10]

(16) ***gamla*** *menn ok* ***unnga*** (homiliebok: 2468)
old men and young
'old and young men'

The N-A-*and*-A pattern is a relatively rare pattern, especially in Old English (see Table 1), and it is difficult to analyze, since the postnominal position is a busy position in terms of different things that can potentially go on there. I will mention one

8 See Fischer (2012) for a careful analysis of the postposed *and*-adjective construction.
9 No data is available for the other *and*-constructions.
10 In addition to 'strict' and 'sloppy', the reference can also be ambiguous.

analysis problem here. If there is a determiner in front of each adjective (N-det+A-*and*-det+A), YCOE analyzes the adjectives as appositions to the noun,[11] whereas Menotec analyzes them as attributes. The apposition analysis is obvious in (17), but not in (18), cf. translations, so this illustrates how compromises sometimes have to be made in corpus annotation. An Old Norwegian example is given in (19).

(17) *ða twa gecyðnyssa þa* ***ealdan*** *and ða* ***niwan***
the two testaments the old and the new
'the two testaments, the old and the new' (cocathom2: 12.1:117.258.2549)

(18) *Oswald, Norðanhymbra cyning se* ***betsta*** *& se* ***cristenesta***
Oswald, Northumbrians' king the best and the most christian
'Oswald, the best and most Christian Northumbrian king' (cobede: 2:5.110.2.1027)

(19) *byrr hinn* ***bazti*** *oc hinn* ***hœgazti*** (strleik: 101)
sailing wind the best and the timeliest
'the best and most timely sailing wind'

What this brief discussion has made clear is that it is impossible to account for the precise distribution of adjectives, especially in the *and*-patterns, without carrying out very detailed queries, combined with manual culling of examples. The cross-linguistic aspect, where data is collected from corpora that are annotated on the basis of different theoretical frameworks, makes it particularly challenging to achieve both good recall and good precision (Ball 1994).

A final point to be mentioned is that Old Norwegian has postnominal possessives, as in (20), whereas postnominal possessives are only used in certain specific constructions in Old English, e.g. *Fæder ure* 'our Father' with reference to God. Differences of this kind may account for the seemingly greater tolerance of Old Norwegian with respect to placing coordinated adjectives postnominally, though it should be kept in mind that the numbers are low in this category.

(20) *misgiærningar varar* ***margar*** *ok* ***mycclar*** (homiliebok: 3495)
misdeeds our numerous and great
'our numerous and great misdeeds'

To sum up, the categories presented in this section give an overview of the distribution, but we have seen that a number of issues should ideally be taken into account, and that it would be possible, and indeed necessary, to create more fine-grained subcategories for each of the patterns in order to fully understand the workings of Old English and Old Norwegian adjectives. We must, however, leave that to future work;

11 They are thus not analyzed as attributive adjectives. There were seven such instances.

we will instead focus on the first of the patterns presented above, namely the A-A-N type, and compare Old English and Old Norwegian. But first we make a detour to Wonderland.

4 Intermezzo: Alice's Adjectives in Wonderland

From time to time a text is translated from the modern language into its earlier version, and so Lewis Carroll's *Alice's Adventures in Wonderland* now exists in Old English under the title *Æðelgýðe Ellendæda on Wundorlande* thanks to the efforts of Peter Baker (2015).

One challenging aspect of this type of translation is of course the vocabulary and how to render modern concepts, often expressed by means of French or Latin loanwords, into Old English with its predominantly Germanic vocabulary. Another one is syntax, since the syntax of English has changed considerably since Old English times. The dilemma for the translator is therefore to what extent the translation should reflect Old English syntax, and to what extent it should be modernized in order to aid the contemporary reader, who may enjoy reading Old English but does not necessarily have much knowledge about actual Old English syntax or the syntactic variation that characterizes this stage of the language.

For the purposes of this intermezzo, I manually extracted all the noun phrases containing two attributive adjectives (A-A-N) from the first five chapters of *Alice*. There were 36 instances. Then I compared these to Baker's translation, to see which strategies he had employed. The results are given in Table 2.

	#	%
A-A-N	15	41.7
A-N-and-A	4	11.1
A-and-A-N	3	8.3
A-N	3	8.3
other	11	30.6
Total	36	100.0

Table 2: The translation of present-day English A-A-N noun phrases into Old English in Baker's *Æðelgýðe Ellendæda on Wundorlande*

As Table 2 shows, Baker often chooses to translate modern A-A-N order into the same order in Old English, cf. (21)–(23).[12]

(21) the **wise little** Alice (p. 13)
séo **wíse lýtle** Æðelgýð (p. 12)

12 The page numbers refer to the pages in the editions used: Carroll (1971) and Baker (2015).

(22) **large round** eyes (p. 37)
miclum sinewealtum éagum (p. 41)

(23) the **distant green** leaves (p. 47)
þám **fyrlenum grénum** léafum (p. 52)

Sometimes Baker coordinates the two prenominal adjectives, as in (24), and sometimes he employs the well-known Old English (and Old Norse) pattern of postponing one of the adjectives, i.e. the A-N-*and*-A pattern, as in (25) (see (10) for authentic examples).

(24) **low trembling** voice (p. 23)
stillre and **bifiendre** stemne (p. 24)

(25) **shrill passionate** voice (p. 22)
sciellre stefne and **grambærr** (p. 23)

It also happens that Baker simply leaves out one of the adjectives, as in (26). All three occurrences of A-N in this little dataset are translations of *the little golden key*. After having introduced the little golden key, Baker, unlike Carroll, apparently does not find it necessary to specify both *little* and *golden* every time.

(26) the **little golden** key (p. 14, p. 17)
þá **lýtlan** cǽge (p. 14)
þá **gyldenan** cǽge (p. 17)

In the category 'other' are found various other strategies, some of which also involve omission, but in addition to something else. In (27), for example, one adjective has been omitted and the remaining adjective is modified by an adverb, perhaps to strengthen the meaning of *ungelæred* so that it corresponds to *ignorant*. In (28) one adjective is omitted and the meaning corresponding to the quantifier *several* occurs postnominally, and in (29) Baker has perhaps reasoned that it is obvious that a kid's hide is white and therefore left out the adjective. In (30) both adjectives are omitted, while in (31) the head noun is omitted and the adjectives are made predicative. In a prepositional expression is used in (32).

In (30) both adjectives are omitted, while in (31) the head noun is omitted and the adjectives are made predicative. In (32) a prepositional expression is used.

(27) an **ignorant little** girl (p. 11)
swíðe ***ungelæred*** *mǽdencild* (p. 10)
very unlearned maid-child

(28) several **nice little** stories (p. 13)
wynsumra *spella* *ná* *féawa* (p. 13)
winsome stories not few

(29) one of the Rabbit's **little white** kid-gloves (p. 19–20)
áne þæs Haran ***lýtelra*** *glófa of ticcenes felle geworhtra* (p. 20)
one of the Rabbit's little gloves of kid's hide made

(30) her eyes immediately met those of a **large blue** caterpillar (p. 39)
þá sóna lócode héo on sumes tréowwyrmes éagan (p. 43)
then immediately looked she into a tree-worm's eyes

(31) She is such a **dear quiet** thing (p. 22)
Héo is swá ***leof*** *and swá* ***smylte*** (p.23)
she is so dear and so quiet

(32) A **little bright-eyed** terrier (p. 23)
Lýtel *eorþhund mid beorhtum éagum* (p.23)
little earth-hound with bright eyes

If we compare Baker's translation to the data from Old English as presented in Table 1, we see that the three most common Old English patterns are also found in Baker's translation of A-A-N order in Alice. However, the difference between authentic Old English and the translation into Old English is that the authentic language rarely allows two descriptive adjectives in an A-A-N pattern (which is the one Baker employs most frequently in his translation), as section 5 will make clear. Hence, without depreciating Baker's impressive achievement in any way, this little exposition reminds us that element order in the early stages of English is not just about syntactic rules, but that more subtle mechanisms, for example to do with semantics, are also at play.

5 The Adjective-Adjective-Noun pattern

We return to the authentic texts. The A-A-N pattern is the most frequent Old English pattern, and the second most frequent Old Norwegian pattern. As regards Old English, Mitchell (1985, §173) comments that "[t]here is room for more work on the arrangements when two attributive adjectives qualify the same noun" without a linking conjunction. He observes that although this pattern seems to be infrequent, it does occur, and hence the claim that Old English adjectives were non-recursive[13] (Spamer 1979, cited in Mitchell 1985 I, §173) does not hold. Fischer (2000, p. 163), however, points out

13 The term 'recursive' is not well defined in any of the sources mentioned here. It seems that the term is used in a wide, non-theoretical sense, referring to adjectives occurring in a series without coordinating conjunction(s).

that Spamer's claim relates to strong adjectives only, since Spamer does not regard weak adjectives as adjectives proper, but as so-called 'adjuncts', i.e. elements which behave like the first part of a compound noun (Fischer 2000, p. 177 fn 8; Spamer 1979, p. 242, 246). Fischer, on the other hand, suggests that neither strong nor weak adjectives are recursive in Old English (2000, p. 171), though she presents some counterexamples for both strong and weak adjectives, which she, interestingly, accounts for in much the same way. She notes that many of them are denominal adjectives referring to material or nationality, or she calls them idiomatic constructions (2000, p. 172–174; see also Fischer 2006, p. 269). Fischer (2001, p. 258) comments that "in Old English adjectives cannot really occur in a row as they do in Present-day English … In Old English two adjectives are either connected by *and* or draped around the noun". In Fischer (2006, p. 253), she says that it was unusual for adjectives to be stacked, whereas Fischer and van der Wurff (2006, p. 125) and Fischer (2012, p. 255 fn 4) say that adjectives could not be stacked. Pysz (2007; 2009, p. 29-34, 208-221) takes both Spamer (1979) and Fischer (2000) to task on empirical grounds, and shows that the number of Old English prenominal stacked adjectives, both weak and strong, is non-negligible.

As we saw in Table 1, there are numerous examples of A-A-N in Old English, and the purpose of this study is to have a closer look at what types of adjectives are found in this construction. My hunch, and hence my hypothesis, was that both Old English and Old Norwegian, unlike present-day English and Norwegian, disallow two descriptive adjectives next to each other.

The next step, then, was to study the distribution of adjectives within this pattern. It immediately became clear that in a majority (186, 62.8%) of the Old English A-A-N constructions, one of the adjectives is *agen* 'own', *ilca* 'same', *oðer* 'other', *self* 'same', or *swilc* 'such'.[14] They are annotated as adjectives because they take adjectival endings, but their degree of 'adjectivity' can be discussed. For our purposes, they can roughly be categorized as peripheral, non-descriptive, determiner-like adjectives, and they easily combine with descriptive adjectives, as in (33) and (34).

(33) *se* ***ylca*** ***arwyrða*** *wer* (cogregdC: 7.49.20.558)
the same honourable man

(34) ***oðrum*** ***langsumum*** *sprǽcum* (coaelive: 86.1263)
other lengthy speeches

Another large group (64, 21.6%) which could be excluded was noun phrases containing classifiers, i.e. adjectives that denote type or origin. In the modern language, such adjectives would typically be found in the prehead position, and are the 'least adjectival and most nominal' of the adjectives (Quirk et al. 1985, p. 1339). Some Old English examples are given in (35)–(38), with the classifiers underlined. The common

14 Fischer (2000, p. 164; 2006, p. 269) also notices these.

adjective *halig* 'holy' was also included in this category, although it can co-occur with a classifier (38).

(35) *se* ***gooda*** ***heofenlica*** *fæder* (cocathhom1: 18.322.150.3542)
the good heavenly father

(36) *þone* ***smyltan*** ***suðanwesternan*** *wind* (coboeth: 4.10.10.122)
the calm southwesterly wind

(37) *anne* ***þicne*** ***linenne*** *clæð* (coherbar: 130.1.1920)
a thick linen cloth

(38) *þære* ***halgan*** ***Romaniscan*** *cirican* (cobede: 2.102.10.961)
the holy Roman church

When these two largest groups had been accounted for, 46 noun phrases remained, so the procedure of evaluating whether the two adjectives were descriptive or not continued. There was a small group (6, 2.0%) of noun phrases with quantifier-like adjectives, as in (39) and (40) (see also Old Norwegian below).

(39) ***mænigfeald*** ***gastlic*** *gewin* (coverhom: 41.1789)
manifold spiritual battle

(40) ***missenlicum*** ***þearfendum*** *mannum* (cogregdC: 28.159.7.1898)
various needy men

The annotation of two of the noun phrases (0.7%) can be discussed, namely (41) and (42). In (41), *unmetlice* is annotated as an adjective in YCOE, but *Bosworth-Toller* considers *unmetlice* to be an adverb in the same example. In Old English *-lic* is an adjective suffix. This adjective suffix can be combined with a case ending, so if *unmetlice* is interpreted as an adjective in (41), *-e* is the nominative plural feminine strong adjectival ending. Adverbs were usually formed from adjectives, e.g. with the suffix *-e* or with the suffix *-lice*. For example, the adjective *freondlic* 'friendly' becomes the adverb *freondlice* 'amicably', and the adjective *blind* 'blind' becomes the adverb *blindlice* 'blindly'. In other words, in Old English *-lice* can either signal adjective + case ending, or an adverb made from an adjective in *-lic*, or an adverb formed with the suffix *-lice*. This means that *unmetlice* in (41) could be interpreted as either an adjective or an adverb. In (42), on the other hand, it is likely that *inlice* is an adverb, since *þing* is a neuter noun and we would therefore not expect a case ending in *-e* for the adjective (accusative here). *Bosworth-Toller* gives the meaning 'thoroughly' in this example, and this fits well with the context.

(41) ***unmetlice*** ***greate*** *heanisse* (coalex: 8.11.45)
immoderate(ly?) great heights

(42) ***inlice*** ***good*** *þing* (coboeth: 34.94.5.1809)
inward(ly?)/thorough(ly?) good thing

We are down to 38 noun phrases, and the most difficult classification remains. When I looked at the remaining examples, it became clear that for many of them there is a hierarchical structure within the noun phrase, such that one adjective has scope over the other. According to Fischer and van der Wurff (2006, p. 125), this is not possible in Old English:

> It seems to be the case that in OE each adjective had the same level with respect to the noun; there was no hierarchy in which one adjective modified the remainder of the NP. It was therefore virtually impossible to put one adjective after another in a row.

I will claim that such a hierarchy is indeed possible in Old English, but it is sometimes difficult to determine it with certainty for specific occurrences. For example, in (43), it is clear that the meaning is that the man is poor (in the emotive meaning) because he is childless (he has lost his son to an evil spirit). In (44), the reference is to a young man who is unknown, not a man who is unknown and young. In (45), the base actions are similar to previous base actions. Hence, the first adjective has scope over the other in these cases.

(43) *se* ***earma*** ***bearnleosa*** *ceorl* (cochdrul: 84.39.1132)
the poor childless churl

(44) *an* ***uncuð*** ***geong*** *man* (cosevens1: 559.438)
an unknown young man

(45) ***gelicum*** ***fullicum*** *weorcum* (comary: 391.254)
similar base actions

At the other end of the scale, we find examples such as (46)–(48), which clearly have two asyndetically coordinated adjectives, i.e. two adjectives that separately describe the noun.

(46) *Cristes* ***soþre*** ***eaþmodlicre*** *andetnesse* (coblick: 171.5.2159)
Christ's true humble confession

(47) *þæt ofstandende* ***þicce*** ***slipige*** *horh* (colaece: 16.1.14.2317)
the remaining thick slimy phlegm

(48) *þa* ***clænan*** ***mildheortan*** *men* (coverhom: 79.2192)
the clean mildhearted men

But then there are some occurrences for which it is difficult to decide whether the adjectives are hierarchically structured or modify the noun independently of each other. Are the tales in (49) false tales that are old, or old and false tales? Are the men in (50) worthy because they are righteous, or worthy and righteous? Is the lust in (51) independently excessive and unclean or not?

(49) ***ealdum*** ***leasum*** *spellum* (coboeth: 35.98.25.1907)
old false tales

(50) *þa* *sawla* *þara* ***fullmedomra*** ***rihtwisra*** *manna* (cogregdC: 26.295.18.4373)
the soul of the worthy righteous men

(51) ***ofermæte*** ***unclæne*** *luste* (comart3: Ja17,A.10.126)
excessive unclean lust

The final count gave 20 instances of noun phrases in which one of the adjectives had scope over the other, and eight of these involved the adjective *earm* 'poor'. I found eight examples which clearly had two descriptive adjectives, and ten examples that were uncertain (among them some duplicates). A summary of the findings for Old English is given in Table 3.

	#	%
One adjective is *agen, ilca, oðer, self,* or *swilc*	186	62.8
Classifiers	64	21.6
Quantifier-like	6	2.0
Possible misannotations with *-lice*	2	0.7
One adjective has scope over the other	20	6.8
Uncertain whether one adjective has scope over the other	10	3.4
Two adjectives describing the noun independently of each other	8	2.7
Total	296	100.0

Table 3: Distribution of adjectives in the Old English A-A-N pattern

We now turn to Old Norwegian. Here, the picture is very clear. In a majority of the A-A-N constructions (27, i.e. 52.9%), the first of the two adjectives is *margr* 'many, numerous', as in (52). The exception is again *Magnus Lagabøte's national law code*, in which most of the constructions involve a numeral annotated as an adjective, e.g. (53).

(52) ***marga*** ***fagra*** *viði* (strleik: 1048)
many beautiful trees

(53) ***xij*** ***skynsamer*** *menn* (mll: 659)
twelve reasonable men

There were also other instances of quantifier-like adjectives, e.g. *margskonar* 'manifold' in (54) and *fyrst* in (55). Altogether, the quantifiers (including *margr*) accounted for 42 of the 51 (82.4%) A-A-N noun phrases in Old Norwegian.

(54) ***margskonar*** ***goðom*** *drycc* (strleik: 1482)
all-kinds-of good drink

(55) ***Fyrst*** ***licamleg*** *synd* (homiliebok: 632)
first bodily sin

Of the remaining nine phrases, six contained a classifier, as in (56) and (57) (see also (6b) above), and in one, the first adjective had scope over the other (58).

(56) ***visum*** ***boc lærðom*** *man{n}um* (homiliebok: 3542)
wise book-learned man

(57) ***dyrlegr*** ***heilagr*** *maðr* (strleik: 1010)
excellent holy man

(58) ***ó orðenna*** ***goðra*** *luta* (homiliebok: 678)
undone (i.e. unperformed) good things

In the end, only two examples remained of what could be termed two stacked descriptive adjectives, namely (59) and (60).

(59) *Sa* *hinn* ***riki*** ***gamle*** *maðr* (strleik: 300)
DET DET rich old man
'the rich old man'

(60) *einum* *hinum* ***bazta*** ***rauðum*** *hesti* (strleik: 1740)
DET DET best red horse
'a splendid red horse'

It seems that for both Old English and Old Norwegian, we can conclude that the A-A-N pattern is dispreferred for noun phrases with two descriptive adjectives, though in order to evaluate it properly, we would need to consider the other patterns in some detail. It might be that two descriptive adjectives are generally uncommon in noun phrases. However, it is likely that when two adjectives are coordinated by means of *and*, as in (9)–(11) above, they are also descriptive.

6 Conclusion

This paper has given an empirical overview of adjective position in Old English and Old Norwegian noun phrases containing two attributive adjectives. The different possible patterns were presented, and one of them, the Adjective-Adjective-Noun pattern, was considered in some detail. The hypothesis was that two stacked adjectives are not both descriptive, and this was borne out. There were a few exceptions, but we would not expect syntactically variable languages like Old English and Old Norwegian to be completely consistent, especially since they have both changed with respect to syntax and word order.

Acknowledgements

I thank two anonymous reviewers and Hildegunn Dirdal for useful comments, Paul Meurer for help in narrowing down the INESS queries for Old Norwegian, Ann Taylor for help with some particularly pesky queries in YCOE, and Helge Dyvik for a discussion of certain points. Search in the Menotec corpus has been made possible through the INESS infrastructure (Rosén et al. 2012), a part of CLARINO (http://clarino.uib.no/iness).

References

Baker, Peter (2015). *Æðelgýðe Ellendæda on Wundorlande*. Portlaoise: Evertype.

Ball, Catherine N. (1994). "Automated Text Analysis: Cautionary Tales." In: *Literary & Linguistic Computing* 9.4, pp. 295–302.

Bosworth-Toller Anglo-Saxon Dictionary (online). URL: http://bosworth.ff.cuni.cz.

Carroll, Lewis (1971). *Alice's Adventures in Wonderland and Through the Looking-glass*. Ed. by Roger Lancelyn Green. London/NewYork/Toronto: Oxford University Press. Originally published by Macmillan (1865 and 1871).

Faarlund, Jan Terje (2004). *The Syntax of Old Norse*. Oxford: Oxford University Press.

Faarlund, Jan Terje, Svein Lie, and Jan Ivar Vannebo (1997). *Norsk referansegrammatkk*. Oslo: Universitetsforlaget.

Fischer, Olga (2000). "The position of the adjective in Old English". In: *Generative Theory and Corpus Studies. A Dialogue from 10 ICEHL*. Ed. by Ricardo Bermúdez-Otero, David Denison, Richard M. Hogg, and C. B. McCully. Berlin/New York: Mouton de Gruyter, pp. 153–181.

– (2001). "The position of the adjective in (old) English from an iconic perspective". In: *The Motivated Sign*. Ed. by Olga Fischer and Max Nänny. Iconicity in Language and Literature 2. Amsterdam: John Benjamins, pp. 249–276.

– (2006). "On the position of adjectives in Middle English." In: *English Language and Linguistics* 10.2, pp. 253–288.

Fischer, Olga (2012). "The status of the postposed '*and*-adjective' construction in Old English: attributive or predicative?" In: *Analysing Older English*. Cambridge: Cambridge University Press, pp. 251–284.

Fischer, Olga and Wim van der Wurff (2006). "Syntax". In: *A History of the English Language*. Ed. by Richard Hogg and David Denison. Cambridge: Cambridge University Press, pp. 109–198.

Haugen, Odd Einar (1995). *Grunnbok i norrønt språk*. Oslo: Gyldendal.

Haumann, Dagmar (2003). "The postnominal '*and* adjective' construction in Old English." In: *English Language and Linguistics* 14.1, pp. 57–83.

– (2010). "Adnominal adjectives in Old English". In: *English Language and Linguistics* 14.1, pp. 53–81.

Lass, Roger (2000). "Language periodization and the concept of 'middle'". In: *Placing Middle English in Context*. Ed. by Irma Taavitsainen, Terttu Nevalainen, Päivi Pahta, and Matti Rissanen. Berlin/New York: Mouton de Gruyter, pp. 7–41.

Mitchell, Bruce (1985). *Old English Syntax*. Vol. 1. Oxford: Oxford University Press.

Nygaard, Marius (1906). *Norrøn syntax*. Kristiania: Aschehoug.

Pysz, Agnieszka (2007). "The (im)possibility of stacking adjectives in Early English". In: *Bells Chiming from the Past: Cultural and Linguistic Studies on Early English*. Ed. by Isabel Moskowich-Spiegel and Begoña Grespo-García. New York/Amsterdam: Rodopi, pp. 15–35.

– (2009). *The Syntax of Prenominal and Postnominal Adjectives in Old English*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. (1985). *A Comprehensive Grammar of the English Language*. London/New York: Longman.

Ringdal, Karl (1918). *Om det attribute adjektivs position i oldnorsk prosa*. Kristiania: Aschehoug.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer, and Helge Dyvik (2012). "An Open Infrastructure for Advanced Treebanking". In: *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*. Ed. by Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco. Istanbul, Turkey, pp. 22–29.

Spamer, James B. (1979). "The development of the definite article in English: A case study of syntactic change". In: *Glossa* 13.2, pp. 241–250.

Taylor, Ann, Anthony Warner, Susan Pintzuk, and Frank Beths (2003). *The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)*. URL: http://www-users.york.ac.uk/~lang22/YcoeHome1.htm.

The Menotec corpus of Old Norwegian (online). URL: http://www.menota.org/menotec.xml.

Valfells, Sigrid and James E. Cathey (1981). *Old Icelandic. An Introductory Course*. Oxford: Oxford University Press.

Judgement, taste and closely related Germanic languages

Robin Cooper

Abstract. We will look at a treatment of the semantics of taste predicates using TTR (Type Theory with Records). The central idea is that we take the notion of judgement from type theory as basic and derive a notion of truth from that, rather than starting from a semantics based on a notion of truth and trying to modify it to include a notion of judgement. Our analysis involves two types of propositions: Austinian propositions, whose components include a situation and a type, and a subtype of Austinian propositions called subjective Austinian propositions, whose components in addition include an agent who makes the judgement that the situation is of the type. We will argue that attitude verbs can select either for propositions in general (subjective or objective) or for subjective propositions, but that there is no type of objective propositions which can be selected for. We will discuss some apparent counterexamples to this from Germanic languages and argue that there is a phenomenon akin to switch reference in certain attitude predicates when their complement involves a subjective proposition.

1 Introduction

The classical view of meaning in model-theoretic semantics is based on the notion of truth in a possible world conceived of as one way the universe could be. While truth is of central importance in semantics, the notion of truth in a possible world characterizing total information about the whole universe seems an unlikely foundation for the kind of natural reasoning that humans engage in as they wander about the actual world gathering partial information about it. Equally unlikely, it seems to me, would be a view that human reasoning is defined in terms of proof-theoretic manipulations of a syntactic calculus somehow implemented in the brain. This does not seem to bode well for explaining how we learn to reason through interaction with the world around us. In this paper we will explore a notion of judgement derived from rich type theoretic approaches to semantics. We will claim that truth is supervenient on judgements

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1333.

that a situation is of a given type. We will argue for this on the basis of important classes of natural language examples where meaning is defined only in terms of subjective judgement and there is no objective truth of the matter, for example *predicates of personal taste* such as *This soup is delicious*, expressions of artistic judgement such as *Stockhausen's* Gruppen *is a masterpiece* and expressions of moral or political judgements such as *Women should be allowed to drive*. An important part of our approach has to do with the kind of reasoning that takes place during interaction in dialogue, and the notion of a dialogue gameboard as proposed by Ginzburg (2012) will play a significant part in the analysis.

2 Predicates of personal taste and other judgements

There is a considerable literature on exchanges involving predicates of personal taste such as (1).

(1) A: This soup tastes great
B: No, it tastes horrible

Among much other work, the following have greatly influenced my own thinking about such examples: Björnsson and Almér (2011), Crespo and Fernández (2011), Lasersohn (2005), and Stojanovic (2007).

What proposition, if any, are A and B disagreeing about? If we adopt the analysis of dialogue proposed by Jonathan Ginzburg, most recently culminating in Ginzburg (2012), the following question arises: What, if anything, gets entered onto A's and B's dialogue gameboards as a commitment resulting from this exchange? A standard approach to these cases is to start from a notion of proposition defined in terms of truth in possible worlds and relativize this notion in some way to context possibly involving A's and B's beliefs.

It seems clear at least that A and B are not agreeing, as shown by (2).

(2) a. A: This soup tastes great
B: #No, I agree, it tastes horrible

b. A: This soup tastes great
B: #You're right, it tastes horrible

We might even go so far as to say that they are disagreeing, at least when we consider the acceptable dialogues in (3).

(3) a. A: This soup tastes great
B: No, I disagree, it tastes horrible

b. A: This soup tastes great
B: ?You're wrong, it tastes horrible

The dialogue (3b) may be somewhat less acceptable, and this may be evidence that disagreement is not always about a simple matter of truth. In the literature on predicates of personal taste this kind of disagreement has been referred to as "Faultless disagreement", that is, disagreeing with somebody on a matter does not necessarily mean that you think they are wrong. There is a clear distinction, for example, between (3b), where how good the soup tastes is intuitively a matter of opinion and (4), where there is intuitively an objective fact of the matter as to whether the soup contains milk.

(4) A: This soup contains milk
B: No, you're wrong, it's coconut milk

When it comes to moral and political judgements and even artistic judgements, issues of right and wrong can arise even if there is perhaps no objective fact of the matter.

(5) a. A: Women should not be allowed to drive
B: No, you're wrong, of course they should

b. A: Stockhausen's *Gruppen* is rubbish
B: No, you're wrong, it's a masterpiece

(5b) is particularly appropriate if B feels that A is ignorant about contemporary music and has no right to venture an uninformed opinion.

3 Strategies for accommodating personal judgements into truth-based semantics

The central question for a truth-based semantics is what A and B in our examples are disagreeing about. Can we find an appropriate proposition that they believe to be true and false respectively? Or can we interpret the personal judgement sentences in a way so that no conflict arises in order to account for the faultless aspect of the disagreement? One way to try to account for the no fault aspect is to say that the personal judgements express attitude reports. This might be realized by saying that (6a) actually expresses (6b).

(6) a. This soup tastes great
b. I think this soup tastes great

Initially, this seems like a plausible idea. However, if the two sentences were equivalent we would accept both of the dialogues in (7) to be equally acceptable

(7) a. A: This soup tastes great
B: ?#You're entitled to your opinion, of course,
but it tastes horrible

b. A: This soup tastes great
B: You're entitled to your opinion, of course,
but I think it tastes horrible

A theory which regards the two sentences in (6) as equivalent would have to explain why they are not substitutable for each other in (7).

A different strategy is to try to exploit the indices that are used for evaluation in a traditional model-theoretic semantics, that is, possible world, time, speaker, hearer, and so on. An obvious place to start is with the speaker of the sentence. We might say that the truth of personal judgements is relativized to the speaker of the sentence. This is schematically represented in (8).

(8) $[\![\text{ This soup tastes great }]\!]^{\ldots,spA,\ldots} \neq$
$[\![\text{ This soup tastes great }]\!]^{\ldots,spB,\ldots}$

This approach would mean that personal judgements are treated as if they contain an implicit first person indexical and so one might expect them to behave similarly to sentences in which there is an overt first person indexical as in (9).

(9) $[\![\text{ I like this soup }]\!]^{\ldots,spA,\ldots} \neq$
$[\![\text{ I like this soup }]\!]^{\ldots,spB,\ldots}$

Unfortunately, however, the sentences with the overt indexical do not at all behave in the same way as personal judgements when it comes to disagreement, as shown in (10).

(10) A: I like this soup
B: #No, I don't /
No, you don't (you're just saying that) / *I* don't

The crucial point here is that you cannot say *No, I don't* in response to *I like this soup* whereas you can say *No, it's not* in response to *This soup is delicious*. Lasersohn (2005) makes a similar point.

It is not even clear that the interpretation of a personal judgement is always relative to the speaker. Consider the examples in (11).

(11) a. Child: This medicine's yucky
Parent: Yes, I know (it's yucky), but it will do you good

b. A: This soup tastes great
B: Does it? (I'm glad / It's horrible /
I can't tell what I think)

There's something more complex than straightforward indexical semantics going on. In (11a) the parent is adopting the perspective of the child. The parent can make this contribution even if the medicine tastes perfectly OK for them as an adult. Similarly (11b) seems to show that a question about a personal judgement most naturally addresses the personal judgement of the hearer, not the speaker. Similarly, the continuation *I'm glad* in this example seems to concern A's judgement of the soup, not B's,

whereas *It's horrible* concerns B's judgement. There seems to be a notion of perspective here similar to the interpretation of spatial expressions like *left* and *right* where we can choose to adopt either our own or another person's perspective. It is different to the spatial case, however, in that with *left* and *right*, once you have fixed a perspective there does appear to be an objective fact of the matter whether one object is to the left or right of another. In the case of taste there does not appear to be a neutral "fact" independent of the perspective. Again Lasersohn (2005) has different examples making a similar point.

4 Judgement and truth

In mainstream semantics truth is central to our explanation of meaning and reasoning. Traditional notions of proposition are based on truth, for example truth in possible worlds. Propositions are regarded as sets of worlds where the proposition is true. In general, the approach to dealing with taste in the literature has been to refine this truth-theoretic approach by adding additional parameters (making truth relative or contextually determined). This has the consequence that ultimately there is some fact of the matter that is *true*, *false* or perhaps *undefined* if we allow truth-value gaps.

In type theory of the kind discussed in Martin-Löf (1984) and Nordström et al. (1990) we get a slightly different spin on this issue. A central notion is that of a *judgement* that an object a is of a type T, $a : T$. We say "a is of type T" or "a is a *witness* for T". There is, of course, a notion of truth in this kind of type theory but it is parasitic on judgement. Types are seen as the truth-bearers (following the dictum of "propositions as types") and types are "true" just in case there is something of the type. This means that types have a dual role: classifying objects and situations on the one hand and serving as truth bearers on the other (corresponding to propositions that there are objects or situations of a given type).

This suggests to us the following strategy for dealing with personal judgements: rather than taking truth as basic and trying to finagle judgement, we take judgement as basic and say that in many cases, though not all, there is, in addition, a fact of the matter. In a general sense, this is a Montagovian strategy: make the apparently more complex case basic and add to it for what you think of as being the ordinary case (cf. Montague's treatment of intensional verbs). Our claim is that we only think of taste predicates as being difficult because we are starting from truth-based semantics rather than judgement-based semantics.

Note that we are *not* saying that truth is not important for semantics. It is still of central importance. Our access to truth in natural (human) reasoning, however, is through judgement and it should not be surprising that this should be reflected in the nature of natural reasoning systems.

5 Type theory and personal judgements

We shall make our proposal in terms of a type theory which we have called TTR, for "Type Theory with Records" (Cooper 2005a; Cooper 2005b; Cooper 2012; Cooper and Ginzburg 2015; Cooper in prep.).

We can think of a judgement as a *type act* in the sense of Cooper (2014). That is, we can be explicit about the role of an agent in the act of judgement: agent A judges object a to be of type T, $a :_A T$. Following a suggestion by Ginzburg (2012) we say that the result of a judgement that a situation s is of type T, $s : T$, can be seen as a type-theoretic object, an *Austinian proposition*, a record with two fields, labelled 'situation' and 'type' as given in (12).

(12)
$$\begin{bmatrix} \text{situation} & = & s \\ \text{type} & = & T \end{bmatrix}$$

The Austinian proposition (12) is true just in case s is indeed of type T. Now let us consider an Austinian proposition where we make the agent explicit. It has an additional field labelled 'agent'.

(13)
$$\begin{bmatrix} \text{situation} & = & s \\ \text{type} & = & T \\ \text{agent} & = & A \end{bmatrix}$$

We call this a *subjective Austinian proposition*. It is true just in case A judges s to be of type T, $s :_A T$. We will call Austinian propositions which have two fields as in (12) *objective Austinian propositions*. As type-theoretic objects these records belong to types. The type *AusProp* of Austinian propositions is (14).

(14)
$$\begin{bmatrix} \text{situation} & : & \mathit{Sit} \\ \text{type} & : & \mathit{Type} \end{bmatrix}$$

A record of this type is required to have two fields labelled 'situation' and 'type' and the objects in those fields must be respectively of type *Sit* ("situation") and *Type* (the type of types[1]). A record with additional fields also belongs to this type. Thus both objective and subjective Austinian propositions are of this type. The type *SubjAusProp* ("subjective Austinian proposition") in addition requires the agent field filled by an object of type *Ind* ("individual"). This is given in (15).

(15)
$$\begin{bmatrix} \text{situation} & : & \mathit{Sit} \\ \text{type} & : & \mathit{Type} \\ \text{agent} & : & \mathit{Ind} \end{bmatrix}$$

1 We avoid Russell's paradox by stratifying the types. See Cooper (in prep. 2012) for discussion.

Thus while the *records* themselves have a fixed finite number of fields, the record *types* do not fully specify how many fields the records of that type should have. The records of a given type must have at least as many fields as specified in the type but they may have more. Thus the records we are using to model objective propositions will have two fields (for a situation and a type), whereas those which model subjective propositions have three fields (that is, with an additional field for the agent). However, the *type* that requires two fields (for a situation and a type) will have witnesses which have exactly the two fields as required, but it will also have witnesses with additional fields. Thus both objective and subjective propositions will be witnesses for the type which requires two fields. This means that the record types introduce a kind of underspecification even though the witnesses for those types are fully determinate with respect to the number of fields that they have.

Record types can be (partially) specified. That is, they can require that a record of the type not only contain appropriate fields with objects of the required type but also that they contain a particular object of the required type. An example is given in (16).

(16)
$$\begin{bmatrix} \text{situation} & : & \mathit{Sit} \\ \text{type=}\mathit{soup\text{-}is\text{-}good} & : & \mathit{Type} \\ \text{agent} & : & \mathit{Ind} \end{bmatrix}$$

Here we have used *soup-is-good* as a representation of the type corresponding to an utterance of *This soup is good*. We are not interested in the exact nature of this type in this paper. (16) is then a partially specified type of subjective propositions.

Our proposal is that in dialogical negotiation we are jointly reasoning about such types of propositions and that these are the objects which are entered into shared commitments on dialogue participants' gameboards, that is, the view of the common ground so far established in the dialogue according to the particular dialogue participant. (One may think of the types as doing the duty of "underspecified representations" of propositions.) Saying *This soup is good* offers the type (17a) or (17b) and claims you can instantiate it with a true proposition.

(17) a.
$$\begin{bmatrix} \text{situation} & : & \mathit{Sit} \\ \text{type=}\mathit{soup\text{-}is\text{-}good} & : & \mathit{Type} \\ \text{agent} & : & \mathit{Ind} \end{bmatrix}$$

b.
$$\begin{bmatrix} \text{situation} & : & \mathit{Sit} \\ \text{type=}\mathit{soup\text{-}is\text{-}good} & : & \mathit{Type} \end{bmatrix}$$

Crucially, we think that it is not determined by the utterance whether the speaker has a subjective or objective proposition in mind and that often subjective opinion is offered or interpreted as objective fact. Answering *yes* (agreeing) means you can

also instantiate it with a true proposition. Answering *no* (disagreeing) means you can instantiate a type with an incompatible type-field (e.g. *soup-is-horrible*).[2]

Gricean dialogue principles govern which individuals you are allowed to instantiate in the agent field in a subjective Austinian proposition. The maxim of quality says that you are only allowed to claim propositions are true if you have evidence. In the simplest case you are only allowed to assert subjective propositions in which you yourself are the agent. However, this flips to the audience in the case of a question such as *Is the soup good?* since the agent giving the answer must obey the maxim of quality and can only instantiate the agent with themselves.

However, this restriction on instantiation only holds in the simplest case. When your dialogue partner has already told you how they feel, then you have evidence for a proposition with them as agent. In such a case you can choose yourself or your dialogue partner as the agent, as illustrated in (18).

(18) A: This medicine tastes yucky
B: No, it doesn't / Yes, I know

Actually, I think both of B's responses are ambiguous as to whether the agent of the judgement is A or B as illustrated by the continuations in (19).

(19) a. No, it doesn't. You're just pretending.
b. No, it doesn't. I think it's delicious.
c. Yes, I know. It's very bitter for young children.
d. Yes, I know. It's dreadfully bitter.

6 Propositional attitudes towards subjective and objective propositions

There are some cases in which it is unclear whether there is an objective fact of the matter or not. Consider (20).

(20) A: (taking a sip of tea) This milk is sour
B: (tasting the milk) No, it's fine

Is there an objective fact concerning whether the milk is sour or not? The dialogue does not seem to force us into a decision. It is therefore fortunate that we have the record type (21), which does not decide the matter.

(21) $\begin{bmatrix} \text{situation} & : & \mathit{Sit} \\ \text{type}=\mathit{milk\text{-}is\text{-}sour} & : & \mathit{Type} \end{bmatrix}$

2 To say that two types are incompatible means that there can be no object which is of both types. See the discussion of negation in Cooper and Ginzburg (2012).

A record, r, is of a given record type, T, just in case r has fields with labels which are the same as the labels in T and objects in those fields which are of the types specified in the respective fields in T. Crucially, r may also have additional fields with labels not in T. Thus an objective proposition not containing an agent could be of the type (21), but so also could a subjective proposition which in addition specifies an agent. Thus (21) does not specify whether the proposition is subjective or not. One suspects that there are many dialogues where it is unspecified as to whether we are dealing with objective facts or not, and this may lead to misunderstanding or even deliberate misinformation. This may have relevance for current political discourse.

It follows from this that any Austinian proposition of the type *SubjAusProp*, as defined in (15), will be of the type *AusProp*, as defined in (14), that is, *SubjAusProp* is a subtype of *AusProp*. The fact that we have a type of Austinian propositions in general and a type of subjective Austinian propositions but not a type of objective Austinian propositions gives us a prediction that predicates of propositional attitudes may select for either of the two types but not for objective propositions as such. In the remainder of this section we will look at some examples and discuss whether this prediction is borne out.

There are clearly verbs which select for subjective propositions but which do not allow objective propositions. Examples with English *find* are given in (22).

(22) a. Anne finds Mary beautiful (Sæbø 2009, p. 336)
b. #Homer finds Bart gay (Sæbø 2009, p. 329)

There are similar examples with German *finden* 'find', as in (23).

(23) a. *Ich finde, dass die Preise hoch sind*
I find that the prices high are
'I find the prices high' (Sæbø 2009, p. 328, modified)
b. *#Die meisten Menschen finden, dass es einen Osterhasen gibt*
the most people find that it a Easter hare gives
'Most people find there to be an Easter Bunny' (Sæbø 2009, p. 328, modified)

Corresponding examples occur with Norwegian *synes* 'think/find', as shown in (24).

(24) a. *Hun synes alle røykere er usympatiske*
she thinks all smokers are unpleasant
'She thinks/finds that all smokers are unpleasant' (Sæbø 2009, p. 339)
b. *#Mange forskere synes at dinosaurene ble utryddet av et*
many researchers thinks that the dinosaurs were exterminated by a
voldsomt kometnedslag
powerful comet strike
'Many researchers find that the dinosaurs were exterminated by a powerful comet strike' (Sæbø 2009, p. 335)

Swedish *tycka* 'think/find' corresponds to Norwegian *synes*, as shown in (25).

(25) a. *Många tycker att kärnkraftverk är vackra*
many think that atomic power stations are beautiful
'Many people find atomic power stations beautiful' (Sæbø 2009, p. 328, modified)

b. *#Många forskare tycker att dinosaurierna* blev *utrotade av ett våldsamt kometnedslag*
many researchers think that the dinosaurs were exterminated by a powerful comet strike
'Many researchers find that the dinosaurs were exterminated by a powerful comet strike'

English *think* is a verb that clearly takes both subjective and objective propositions as complement, as shown in (26).

(26) a. Many people think that atomic power stations are beautiful

b. Many researchers think that the dinosaurs became extinct because of a gigantic comet striking the earth

Potential counterexamples for our prediction are those which appear to take only objective propositions as complements. A candidate is English *believe* as in (27).

(27) a. Many researchers believe that the dinosaurs became extinct because of a gigantic comet striking the earth

b. ?Many people believe this soup is delicious

Swedish *tro* 'believe' seems similar to English *believe* as shown in (28).

(28) a. *Många forskare tror att dinosaurierna blev utrotade av ett våldsamt kometnedslag*
many researchers believe that the dinosaurs were exterminated by a powerful comet strike
'Many researchers believe that the dinosaurs were exterminated by a powerful comet strike'

b. ?*Många tror att denna soppa smakar utmärkt*
many believe that this soup tastes excellent
'Many people believe that this soup tastes excellent'

However, this data concerning *believe/tro* is a little misleading. While they appear to select for objective propositions they can in fact also occur with subjective propositions. It is just that when they do so, they induce an effect which is a little similar to

switch reference: they require that the agent making the judgement in the subjective proposition in the complement is distinct from the subject of the embedding attitude verb. Thus (29) is something you can say when you have not yet tasted the soup but you have heard from somebody else that they thought it was good.

(29) I believe the soup is good

There are similar examples with Swedish *tro* as shown in (30).

(30) a. *Vårt minne lurar oss att tro att människor med ett attraktivt yttre*
our memory tricks us to believe that people with an attractive exterior
är trevligare än fula
are nicer than ugly
'Our memory tricks us into believing that people with an attractive exterior are nicer than ugly people' (http://www.suntliv.nu/Amnen/Halsa/Artiklar-om-halsa/Darfor-tror-vi-att-vackra-manniskor-ar-trevligare/, retrieved Oct. 29, 2012)

b. *Jag önskar att jag vore lika vacker som Sebastian tror att han*
I wish that I were as good looking as Sebastian believes that he
är
is
'I wish I were as good looking as Sebastian believes he is' (title of a novel by Christer Hermansson, 2003)

c. *trots den pinsamma missen tror jag soppan var rätt så god*
in spite of the embarassing mistake believe I the soup was right so good
ändå
anyway
'in spite of the embarassing mistake I believe the soup was pretty good anyway'
(http://tantgulsblogg.se/en-spicy-thai-soppa-och-tant-gul-tokar-till-det/, retrieved Oct 29, 2012)

In (30a) we believe that attractive people are generally judged to be nicer, not just that we think they are nicer. In (30b) Sebastian believes that other people find him good-looking. Finally, in (30c) it is believed that the people to whom the soup was served judged it to taste good. In (31) we give a couple of constructed minimal pairs[3] which further illustrate the point for Swedish.

(31) a. *Jag tror att medicinen smakar gott*
I believe that the medicine tastes good
'I believe the medicine (will) taste(s) good'

3 These are based on examples offered by one of the reviewers.

b. *Jag tycker att medicinen smakar gott*
I think that the medicine tastes good
'I think the medicine tastes good'

c. *Kim tror att små hundar är attraktiva*
Kim believes that small dogs are attractive
'Kim believes that small dogs are attractive'

d. *Kim tycker att små hundar är attraktiva*
Kim thinks that small dogs are attractive
'Kim thinks that small dogs are attractive'

(31a) is something you might say to a child in order to persuade them to take the medicine. You may not have tasted the medicine yourself but you think the child will like it. (31b), on the other hand, means that you have tasted it yourself and your opinion is that it tastes good. You may also use this to persuade the child to take the medicine but here the argument would be "I think it's good, therefore you will think it's good". The sentence only expresses the antecedent in this argument. (31c) can be used to express that Kim believes that people find small dogs attractive, that is, people in general, possibly including Kim, though not necessarily. Think of a situation where Kim is opening a pet shop and wondering what animals to sell — Kim may not personally find small dogs attractive. (31d), on the other hand, requires that Kim personally finds small dogs attractive and cannot be used to say anything about the view of people in general (beyond, possibly, the unexpressed suggestion that if Kim finds small dogs attractive then other people will too).

If English *believe/think* make a similar distinction in terms of subjectivity with respect to their complement as Swedish *tro/tycka*, why do the latter seem so exotic and difficult for English speakers of Swedish? I believe that the reason for this is that the two pairs divide up the space of possibilities slightly differently. In Table 1, '+' means that the verb takes a subjective/objective complement, '–' means that it does not, and 'sr' means that the verb has the switch reference-like effect on the complement discussed above.

	subjective	objective
believe	sr	+
think	+	+

	subjective	objective
tro	sr	+
tycka	+	–

Table 1: The space of possibilities for the English and Swedish verbs

The fact that *tycka* does not take objective propositions as complement whereas *think* does is apparently enough to cause confusion among English non-native speakers of Swedish.

7 Conclusion

We have argued that natural human reasoning is judgement based rather than truth based and that truth is parasitic on judgement. This seems a reasonable conclusion for agents whose access to truth is through judgement (their own or somebody else's). By this we do not wish to say that there is no notion of objective truth in natural reasoning. We suggest, however, that there are subjective judgements illustrated in natural language by a variety of personal judgements involving taste, morality and artistic judgement among other things, for which there is no objective fact of the matter. As natural reasoners, humans seem deeply engaged in discussing such judgements. They are simply interested in the judgements that other people make.

The distinction between subjective and objective propositions in our particular type-theoretic formulation results in two types of propositions: propositions in general which may or may not specify an agent making the judgement, and a subtype of this type for subjective propositions which do require an agent as the judge. We test the hypothesis that these two types can be selected for by predicates of propositional attitude, but that there is no predicate which can select for only objective propositions (since no such type is available according to our analysis). We have argued against some expected counterexamples to this claim.

Acknowledgements

This paper is for Helge Dyvik, a scholar of judgement, taste and closely related Germanic languages.

I am grateful to Elisabet Engdahl for discussion and to two anonymous reviewers for helpful comments. This research was supported in part by VR project 2009-1569, Semantic analysis of interaction and coordination in dialogue (SAICD) and VR project 2013-4873, Networks and Types (Nettypes). Previous versions of parts of this paper have been presented at the workshop Dialogue with Contextualism, Université Paris-Diderot, Sept. 18, 2012; at King's College, London, Oct. 18, 2012; at Trinity College Dublin, Feb. 15, 2013; at the 25th Scandinavian Conference of Linguistics, Reykjavík, May, 13–15, 2013; and at the Conference on Computing Natural Reasoning (CoCoNat'15), July 19–20, 2015, Indiana University, Bloomington.

References

Björnsson, Gunnar and Alexander Almér (2011). "The Pragmatics of Insensitive Assessments". In: *The Baltic International Yearbook of Cognition, Logic and Communication* 6, pp. 1–45.

Cooper, Robin (in prep.). "Type theory and language: from perception to linguistic communication". Draft of book chapters available from https://sites.google.com/site/typetheorywithrecords/drafts.

Cooper, Robin (2005a). "Austinian truth, attitudes and type theory". In: *Research on Language and Computation* 3, pp. 333–362.
– (2005b). "Records and Record Types in Semantic Theory". In: *Journal of Logic and Computation* 15.2, pp. 99–112.
– (2012). "Type Theory and Semantics in Flux". In: *Handbook of the Philosophy of Science*. Ed. by Ruth Kempson, Nicholas Asher, and Tim Fernando. Vol. 14: Philosophy of Linguistics. General editors: Dov M. Gabbay, Paul Thagard and John Woods. Elsevier BV, pp. 271–323.
– (2014). "How to do things with types". In: *Joint Proceedings of the Second Workshop on Natural Language and Computer Science (NLCS 2014) & 1st International Workshop on Natural Language Services for Reasoners (NLSR 2014) July 17-18, 2014 Vienna, Austria*. Ed. by Valeria de Paiva, Walther Neuper, Pedro Quaresma, Christian Retoré, Lawrence S. Moss, and Jordi Saludes. Center for Informatics and Systems of the University of Coimbra, pp. 149–158.
Cooper, Robin and Jonathan Ginzburg (2012). "Negative inquisitiveness and alternatives-based negation". In: *Logic, Language and Meaning: 18th Amsterdam Colloquium, Amsterdam, The Netherlands, December 19–21, 2011, Revised Selected Papers*. Ed. by Maria Aloni, Vadim Kimmelman, Floris Roelofsen, Galit W. Sassoon, Katrin Schulz, and Matthijs Westera. Lecture Notes in Computer Science 7218. Springer, pp. 32–41.
– (2015). "Type Theory with Records for Natural Language Semantics". In: *The Handbook of Contemporary Semantic Theory*. Ed. by Shalom Lappin and Chris Fox. second. Wiley-Blackwell, pp. 375–407.
Crespo, Inés and Raquel Fernández (2011). "Expressing Taste in Dialogue". In: *SemDial 2011 (Los Angelogue): Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue*. Ed. by Ron Artstein, Mark Core, David DeVault, Kallirroi Georgila, Elsi Kaiser, and Amanda Stent, pp. 84–93.
Ginzburg, Jonathan (2012). *The Interactive Stance: Meaning for Conversation*. Oxford: Oxford University Press.
Lasersohn, Peter (2005). "Context Dependence, Disagreement, and Predicates of Personal Taste". In: *Linguistics and Philosophy* 28, pp. 643–686.
Martin-Löf, Per (1984). *Intuitionistic Type Theory*. Naples: Bibliopolis.
Nordström, Bengt, Kent Petersson, and Jan M. Smith (1990). *Programming in Martin-Löf's Type Theory*. Vol. 7. International Series of Monographs on Computer Science. Oxford: Clarendon Press.
Sæbø, Kjell Johan (2009). "Judgment ascription". In: *Linguistics and Philosophy* 32, pp. 327–352. DOI: 10.1007/s10988-009-9063-4.
Stojanovic, Isidora (2007). "Talking about taste: disagreement, implicit arguments, and relative truth". In: *Linguistics and Philosophy* 30.6, pp. 691–706.

Unlike phrase structure category coordination

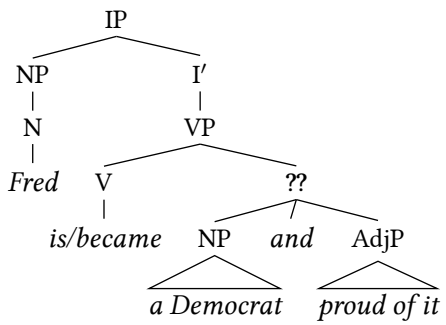
Mary Dalrymple

Abstract. We present a feature-based theory of phrase structure category labels which assigns an appropriate category to unlike category coordinations such as *(Fred is) [[a Democrat] and [proud of it]]*. We propose that unlike category coordinations are specified as including features of the phrase structure categories of each of the conjuncts.

1 Introduction

Often, some syntactic or semantic property of a coordinate structure depends on the corresponding properties of its conjuncts. In this paper we address a particular aspect of that phenomenon: determining the c-structure category label of a coordinate structure in which the conjuncts have different categories. For example, what is the category label of *a Democrat and proud of it* in an example like (1)?

(1)



The c-structure category of a phrase is relevant for category selection requirements imposed by certain predicates and certain phrase structure configurations. These requirements must also be satisfied by coordinate structures, including unlike category coordination. We provide an analysis which assumes that tree nodes are labeled by sets of features, and we propose a means for determining the set of features defining the label of a coordinate structure on the basis of the features labeling the conjunct phrases.

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1332.

2 Category selection by predicate or rule

2.1 Predicates selecting c-structure category

There are very few predicates that require a particular c-structure category for their arguments, but a few such predicates are attested. Often, the verbs *wax* and *become* are given as examples.

As discussed in detail by Pollard and Sag (1994), *wax* in its predicative use requires an adjective phrase complement, and disallows nominal, verbal, prepositional, and adverbial phrase complements.

(2) *a.* Fred waxed [poetical]$_{\text{AdjP}}$/[lyrical]$_{\text{AdjP}}$.
b. *Fred waxed [a success]$_{\text{NP}}$.
c. *Fred waxed [in a good mood]$_{\text{PP}}$.
d. *Fred waxed [waving his arms wildly]$_{\text{VP}}$.
e. *Fred waxed [quickly]$_{\text{AdvP}}$.

Pollard and Sag (1994) claim that *become* requires either a nominal or an adjectival complement.

(3) *a.* Fred became [happy]$_{\text{AdjP}}$.
b. Fred became [a professor]$_{\text{NP}}$.
c. *Fred became [in the room]$_{\text{PP}}$.
d. *Fred became [waving his arms wildly]$_{\text{VP}}$.
e. *Fred became [happily]$_{\text{AdvP}}$.

Be can take an adjectival, a nominal, or a prepositional complement.

(4) *a.* Fred is [happy]$_{\text{AdjP}}$.
b. Fred is [a professor]$_{\text{NP}}$.
c. Fred is [in the room]$_{\text{PP}}$.

In coordination, these requirements are preserved. The complement of *wax* can be a coordinate structure composed of adjective phrases, but no other categories. *Become* allows a coordinate structure composed of an adjective phrase and a nominal phrase, but other categories are not allowed. *Be* allows any combination of adjectival, nominal, and prepositional phrase conjuncts. These constraints are exemplified in (5)–(7), including naturally occurring corpus examples from Wikipedia (Davies 2015).

(5) *a.* Fred waxed [poetical]$_{\text{AdjP}}$ and [philosophical]$_{\text{AdjP}}$.

b. *Fred waxed [poetical]$_{\text{AdjP}}$ and [waving his arms wildly]$_{\text{VP}}$.

(6) *a.* Fred became [a professor]$_{\text{NP}}$ and [proud of his work]$_{\text{AdjP}}$.

b. Some Biblical minimalists like Thomas L. Thompson go further, arguing that Jerusalem became [a city]$_{\text{NP}}$ and [capable of being a state capital]$_{\text{AdjP}}$ only in the mid-7th century. (Wikipedia)

c. *Fred became [a professor]$_{\text{NP}}$ and [in line for a promotion]$_{\text{PP}}$.

d. *Fred became [a professor]$_{\text{NP}}$ and [waving his arms wildly]$_{\text{VP}}$.

(7) *a.* Fred is [a professor]$_{\text{NP}}$ and [proud of his work]$_{\text{AdjP}}$.

b. She accepts her status as a Muggle-born witch, and states in Deathly Hallows that she is "[a Mudblood]$_{\text{NP}}$ and [proud of it]$_{\text{AdjP}}$". (Wikipedia)

c. Fred is [a professor]$_{\text{NP}}$ and [in a good mood]$_{\text{PP}}$.

d. Divion is [a commune]$_{\text{NP}}$ and [in the Pas-de-Calais department in the Nord-Pas-de-Calais region of France]$_{\text{PP}}$. (Wikipedia)

e. Fred is [proud of his work]$_{\text{AdjP}}$ and [in a good mood]$_{\text{PP}}$.

f. Cassie discovers weeks later that the doctor who performed her procedure has been influenced by Azazeal, and that the baby is [alive]$_{\text{AdjP}}$ and [in Azazeal's care]$_{\text{PP}}$. (Wikipedia)

2.2 Phrase structure requirements

C-structure category requirements are not imposed only by predicates on their arguments; c-structure positions can also be restricted to phrases of particular types. For example, the complement position of an English PP can be filled by NP or PP, but not CP.[1]

(8) *a.* I removed it from [the box]$_{\text{NP}}$.

b. I removed it from [under the bed]$_{\text{PP}}$.

c. *I didn't care about [that he might be unhappy]$_{\text{CP}}$.

However, the proper generalization governing these examples concerns the permitted categories of phrases appearing in the complement position of PP, and not the category of the f-structure object of a preposition. In fact, a CP can be the f-structure object of a preposition if it is displaced, as discussed by Kaplan and Bresnan (1982), Kaplan and Zaenen (1989), and Dalrymple and Lødrup (2000).

1 Evidence that the PP *under the bed* is the object of *from*, and that *from under* is not a complex preposition, includes the possibility that *under the bed* can be clefted as in (a), and modified as in (b):

(a) It was [under the bed] that I removed it from.
(b) I removed it from [right/directly under the bed].

(9) [That he might be unhappy]$_{CP}$, I didn't care about.

This means that we cannot rule out examples like (8c) by restricting the category of the f-structure object of the preposition, since this would incorrectly rule out examples such as (9), in which a displaced CP is the f-structure object of the preposition *about*. Instead, we must restrict the phrasal category that can appear as the complement of P in the P′ rule.

Notably, a prepositional complement can also be a coordinate structure with one NP conjunct and one PP conjunct; example (10b) is from the NOW corpus (Davies 2013).

(10) *a.* I removed them from [the box]$_{NP}$ and [under the bed]$_{PP}$.

b. Every year, the Canadian Tourism Commission invites travel journalists from [this country]$_{NP}$ and [around the world]$_{PP}$ to a convention called GoMedia to meet tourism representatives from across Canada. (NOW Corpus)

A c-structure rule allowing a disjunction of NP and PP as the complement of P allows either (conjoined) NPs or (conjoined) PPs, but fails to allow unlike category coordination structures such as *[the box]*$_{NP}$ *and [under the bed]*$_{PP}$, with one NP conjunct and one PP conjunct.

(11) P′ rule, version 1 (unsuccessful):

$$P' \longrightarrow P \quad \{NP \mid PP\}$$

In sum, these examples show the need for a theory of phrase structure category labels that provides an appropriate label for a coordinate phrase composed of unlike categories. A coordination of like categories should have that category, and a coordination of unlike categories should have properties of both categories, or be indeterminate between the two categories in some sense.

3 Previous work and alternative analyses

3.1 Ellipsis?

Beavers and Sag (2004) propose that examples like (12) do not exemplify unlike category coordination, but are in fact coordinated verb phrases with elision of the verb in the second conjunct.

(12) Fred [is a professor] and [~~is~~ proud of his work].

On this analysis, *a professor and proud of his work* is not a constituent, since the second conjunct is analyzed as a subpart of an elided larger structure. Although this is a possible analysis of some examples of this type, it does not constitute a general solution to the problem of unlike category coordination. Peterson (2004) provides two arguments

that an unlike category coordination must have an analysis as a single constituent. First, fronting is possible only for single constituents (13a), but an unlike category coordination can be fronted (13b,c).

(13) *a.* *[A book] [to Fred] though Bill gave…

b. [[A plumber] and [making a fortune]] though Bill may be, he's not going to be invited to my party.

c. [[In town] and [itching for a fight]] is the scourge of the West, Zitty Zeke.

However, this argument is not conclusive: as pointed out by Beavers and Sag (2004), there is an alternative analysis of these examples that conforms to their ellipsis-based approach.

(14) A plumber ~~though Bill may be~~ and making a fortune though Bill may be, he's not going to be invited to my party.

Peterson (2004) provides a second argument based on right node raising, which for at least some English speakers is possible only for single constituents (Bresnan 1974). For those speakers, the examples in (15) show that coordinated unlike categories can form a single constituent, and cannot be analyzed in terms of ellipsis.

(15) *a.* Bill is, and John soon will be, [[a master plumber] and [making a fortune]].

b. I can picture Zeke, but cannot imagine John, [[a convicted felon] and [imprisoned for life]].

An additional difficulty comes from the acceptability of modifiers such as *simultaneously* or *alternately*, for which an ellipsis-based analysis does not produce the right reading; under an ellipsis-based analysis, the examples in (16a) and (17a) are elided versions of (16b) and (17b), but the meanings of the (a) and (b) examples are not the same.

(16) *a.* Fred is simultaneously [a professor] and [ashamed of his work].

b. Fred [is simultaneously a professor] and [~~is simultaneously~~ ashamed of his work].

(17) *a.* Fred is alternately [in a good mood] and [suicidal].

b. Fred [is alternately in a good mood] and [~~is alternately~~ suicidal].

3.2 Choose a new category?

Patejuk (2015) proposes that all unlike category coordination structures have the same c-structure category label; she proposes XP (or, alternatively, UP), which is not a variable over category labels, but a special label for unlike category coordinations. In other words, all unlike category coordinations have the category XP.

(18) XP $\longrightarrow$ YP Conj ZP

This proposal requires potentially radical modification of the grammar to allow the special category XP as well as standard categories like NP and PP wherever unlike category coordination structures can appear. Even when this is done, the proposal does not allow the possibility of imposing the category requirements that were shown to be necessary in Section 2, since on this view all unlike category coordinations have the same category. It also makes it difficult to enforce category-function correlations and to control the distribution of phrases of different categories, since there is no relation between the category of the unlike category coordination structure and the categories of the conjuncts.

3.3 Choose one of the categories?

Peterson (2004) proposes that the category of the coordinate structure in unlike category coordination is the category of the first daughter.

(19) X $\longrightarrow$ X Conj Y

This analysis makes the incorrect prediction that the distribution of an unlike category coordination structure matches the distribution of the category of the first conjunct. As shown in examples (20) and (21), both conjuncts must satisfy the requirements, not just the first one.

(20) *a.* Fred waxed [poetical]$_{\text{AdjP}}$ and [philosophical]$_{\text{AdjP}}$.
b. *Fred waxed [poetical]$_{\text{AdjP}}$ and [waving his arms wildly]$_{\text{VP}}$.

(21) *a.* Fred became [a professor]$_{\text{NP}}$ and [happy]$_{\text{AdjP}}$.
b. *Fred became [a professor]$_{\text{NP}}$ and [in line for a promotion]$_{\text{PP}}$.

In fact, the problem is more general: this proposal allows an unlike category coordination structure to appear wherever the category of the initial conjunct is allowed. For example, if the grammar allows the category CP as a verbal complement in V′, this proposal predicts that any unlike category coordination structure whose first conjunct is a CP is also an acceptable verbal complement, no matter what the categories of the non-initial conjuncts are. Like the Patejuk proposal, then, the Peterson proposal does not enforce category-function correlations or allow for control over the distribution of phrases of different categories, since unlike category coordination structures can contain non-initial conjuncts of any category.

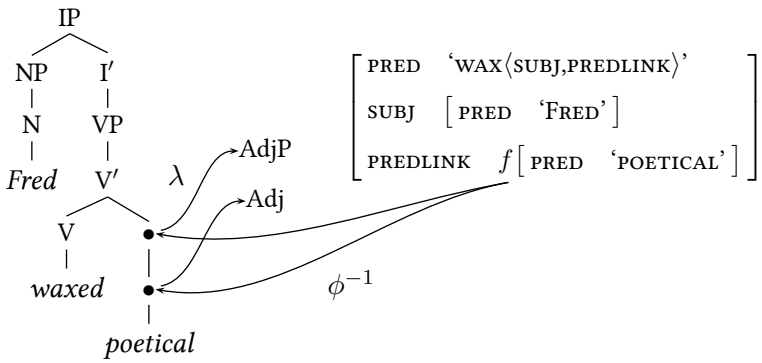
4 Category selection: the CAT predicate

In LFG, category selection by a predicate is treated by appeal to the CAT predicate, which is defined in terms of the node labeling function λ. Nodes in a tree are generally represented by their labels, as in the tree on the left in (22), but this is in fact an abbreviatory convention for the representation on the right, where the λ node labeling function is made explicit.

(22) Standard representation: Making the λ node labeling function explicit:

We can refer to the nodes corresponding to a particular f-structure through the inverse ϕ correspondence: ϕ is a function from c-structure nodes to f-structures, and its inverse ϕ^{-1} is a relation between f-structures and the c-structure nodes that correspond to them.

(23)



This allows us to define the CAT predicate, which relates an f-structure to the labels of the c-structure nodes that correspond to it.

(24) Definition of CAT (Crouch et al. 2008; Kaplan and Maxwell 1996):

CAT(f, C) iff $\exists n \in \phi^{-1}(f) : \lambda(n) \in C$.

"CAT(f, C) is true if and only if there is some node n that corresponds to f via the inverse ϕ correspondence (ϕ^{-1}) whose label (λ) is in the set of categories C."

The CAT predicate allows us to constrain the category of the complement of the verb *wax* by requiring AdjP to be one of the categories of the c-structure nodes correspond-

ing to the PREDLINK of *wax*.[2] The lexical entry for the predicate *wax* using the definition of CAT in (24) is given in (25).

(25) CAT((↑ PREDLINK), {AdjP})

"The label AdjP must be a member of the set of labels of c-structure nodes corresponding to my PREDLINK."

The lexical entry for *become*, which requires AdjP or NP, is given in (26).

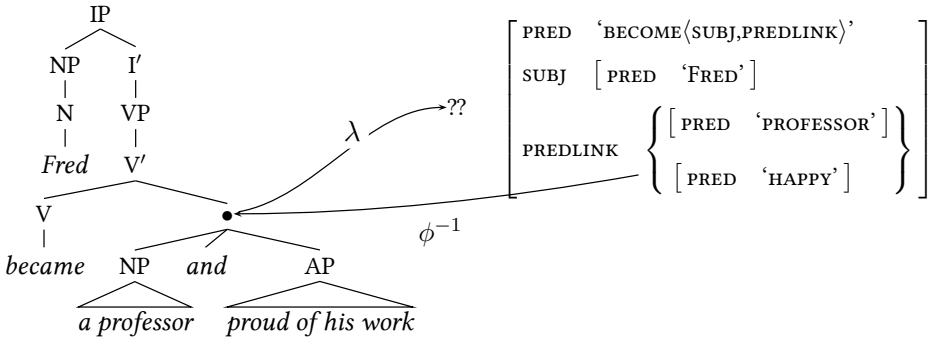
(26) CAT((↑ PREDLINK), {AdjP, NP})

"Either the label AdjP or the label NP must be a member of the set of labels of c-structure nodes corresponding to my PREDLINK."

4.1 CAT in unlike category coordination

What predictions does the CAT predicate make for unlike category coordinations, as in (27)? Notice that the c-structure constituent corresponding to the PREDLINK of *become* is the unlike category coordinate structure *a professor and proud of his work*, as shown in (27), with '??' as the as-yet undefined label for the unlike category coordinate structure.

(27)



This means that we need some additional assumptions to appropriately constrain the categories of the conjuncts in unlike category coordination.

Kaplan and Maxwell (1996) and Crouch et al. (2008) address this problem by proposing that the CAT predicate is *distributive* (Dalrymple and Kaplan 2000): if the CAT predicate is applied to a set of f-structures, it must hold for each member of the set.

2 We analyze the predicative complement of *wax* as the closed grammatical function PREDLINK (Butt et al. 1999; Dalrymple, Dyvik, and King 2004) rather than the open complement XCOMP, but nothing in our analysis hinges on this choice.

(28) For any distributive property P and set s, $P(s)$ iff $\forall f \in s.P(f)$. (Dalrymple and Kaplan 2000, example (73))

Treating CAT as distributive means that each conjunct of a coordinate structure, including unlike category coordinations, must satisfy the constraints imposed by the CAT predicate. If we assume the CAT constraint given in (26) for *become*, the result is as desired for example (27): each conjunct of the coordinated PREDLINK has either the label AdjP or the label NP. Thus, in the analysis of category constraints imposed by predicates such as *wax* and *become* (Section 2.1), treating CAT as distributive yields similar empirical coverage to the solution we propose here.

However, treating CAT as distributive leaves open the issue that is the main focus of this paper: the category label of unlike category coordination structures. Proponents of the distributive CAT definition generally assume the Peterson proposal outlined in Section 3.3, that the category of a nonconstituent coordinate structure is the same as the category of the initial conjunct (Ron Kaplan, p.c.). We must then reevaluate the problem of category selection that arises for the Peterson proposal: recall that the rule in (11) does not adequately constrain the P′ rule, since it allows unlike category coordination structures with an NP or PP initial conjunct and non-initial conjuncts of other categories. This problem can in fact be addressed if the rule is formulated as in (29), with explicit CAT constraints to ensure that all conjuncts in a coordinated prepositional complement are either NP or PP.

(29)
$$\begin{array}{lcl} \text{P}' & \longrightarrow & \begin{array}{ccccc} \text{P} & \{ & \text{NP} & | & \text{PP} \\ \uparrow=\downarrow & & (\uparrow \textsc{obj})=\downarrow & & (\uparrow \textsc{obj})=\downarrow \\ & & \text{CAT}(\downarrow, \{\text{NP PP}\}) & & \text{CAT}(\downarrow, \{\text{NP PP}\}) \end{array} \quad \} \end{array}$$

In fact, such annotations would have to appear throughout the grammar, to prevent the appearance of unlike category coordination structures with conjuncts that are not allowed in particular contexts. For example, to ensure that only phrases of category CP or conjunctions with CP conjuncts can bear the f-structure role of COMP, the CAT annotation in (30) is necessary.

(30)
$$\begin{array}{lcl} \text{V}' & \longrightarrow & \text{V} \quad \begin{array}{c} \text{CP} \\ (\uparrow \textsc{comp})=\downarrow \\ \text{CAT}(\downarrow, \{\text{CP}\}) \end{array} \end{array}$$

We prefer a solution which does not require such a proliferation of additional category constraints. The solution we propose in the following assigns a c-structure category to coordinate structures which reflects the categories of the conjuncts, with an overspecified category reflecting the categories of the conjuncts in unlike category coordination. In this setting, the CAT predicate constrains the category of the coordinate structure as a whole: we advocate a nondistributive definition of CAT which does not distribute to the members of a set.

4.2 Overspecification and indeterminacy

Our analysis is prefigured in work within the GPSG and HPSG frameworks by Gazdar et al. (1985), Sag et al. (1985), and Sag (2002), who propose that there is a systematic relation between the features of a coordinate structure (including features defining the category label as well as other grammatical features) and the features of the conjuncts. Our analysis is also clearly related to work by Bayer (1996), who adopts a deductive approach in a Categorial Grammar setting, and proposes overspecified categories for unlike category coordination.

Gazdar et al. (1985) and Sag et al. (1985) propose that the features of a coordinate structure are the intersection (or, equivalently, the generalization) of the features of the conjuncts, and that a predicate can impose underspecified requirements on its arguments. On their theory, the conjuncts in an example like *[a sick man]*$_{[+\text{N}, -\text{V}, +\text{PRED}]}$ *and [suffering from fever]*$_{[-\text{N}, +\text{V}, +\text{PRED}]}$ have the features indicated. The values of the N and V features clash, but the +PRED feature is common to both conjuncts, and so the coordinate structure has only the feature +PRED. A verb like *is* places no constraints on the N and V features of its complement, requiring only the +PRED feature, and so *a sick man and suffering from fever* is correctly predicted to be an acceptable complement for *is*. Jacobson (1987) and Sag (2002) point out some problems for this proposal when predicates are coordinated: for example, if the predicate *grew* requires a AdjP complement bearing the features [+N, +V] and the predicate *remained* requires an AdjP or NP complement bearing only the feature [+N], Gazdar et al. (1985) and Sag et al. (1985) predict that the coordinated predicates *grew and remained* require only [+N], incorrectly classifying **Kim grew and remained a Republican* as grammatical. Though it does not suffer from these difficulties, our proposal is similar to the Gazdar et al. (1985) and Sag et al. (1985) approaches in that it allows a predicate to place underspecified requirements on the category of its argument: a verb like *become* places indeterminate requirements on its complement, allowing a noun phrase, an adjective phrase, or a coordinate structure with one or more NP conjuncts and one or more AdjP conjuncts.

Sag (2002) proposes a treatment of coordinate structures which allows underspecification in the type lattice, treating only a subset of grammatical features (crucially not including subcategorization requirements) via underspecification in coordination. Bayer (1996) provides an analysis which is similar to Sag's analysis in some respects, according to which unlike category coordinations have a disjunctively specified category label; for example, Bayer proposes the category NP∨S for an unlike coordinate phrase containing an NP conjunct and an S conjunct. Some predicates place fully specified category requirements on their argument; for example, a predicate such as *wax* requires a complement of category AdjP. Other predicates impose a disjunctive category requirement; for example, a predicate such as *become* requires an argument of category NP∨AdjP. A noun phrase such as *a man* is of category NP, but its category can be weakened to NP∨AdjP, allowing it to serve as the complement of *become*. An un-

like category coordination necessarily has a disjunctive category specification, which cannot be strengthened by eliminating one of the disjuncts; for this reason, an unlike NP∨AdjP coordination cannot serve as the complement to a verb like *wax*, which requires the stronger, nondisjunctive category AdjP. Our proposal resembles Sag's and Bayer's in that an unlike category coordination is specified as belonging to each of the categories of the conjuncts, and can appear only with a predicate which places indeterminate category requirements on its argument.

One important difference between these works and our proposal relates to the modular architecture of LFG and the separation of c-structure and f-structure. Our analysis does not assume that f-structure features such as case, person, or number must be treated by the same rules and processes as c-structure features defining category labels. Although our analysis of unlike category coordination bears important similarities to the analysis of f-structure feature indeterminacy, there are also important differences. For example, in the treatment of case indeterminacy (described in the next section) coordinate **predicates** place possibly overspecified requirements on the case features of their shared **arguments**, while arguments can be indeterminately specified, using negative features to rule out unacceptable possibilities. In contrast, in unlike category coordination it is coordinated **arguments** that are potentially overspecified for phrase structure category features, while predicates place potentially indeterminate category requirements on their arguments.

5 Background: F-structure indeterminacy and overspecification

We treat the category of an unlike category coordination as overspecified: that is, an unlike category coordinate structure is specified as belonging to each of the phrase structure categories of its conjuncts. When the categories of all of the conjuncts in a coordinate structure are compatible with the (possibly underspecified) requirements of the governing predicate and the phrasal configuration, an unlike category coordination is acceptable. Our analysis is similar to the Dalrymple, King, et al. (2009) analysis of f-structure indeterminacy, building on the set-based treatment of Dalrymple and Kaplan (2000), which we now describe.

The masculine weak declension plural German noun *Papageien* 'parrots', which shows no case distinctions, can satisfy different case requirements, occurring with verbs that take accusative objects (31) as well as with those that take dative objects (32).

(31) *a.* *Er* *findet* *ihn/*ihm.*
he finds him[ACC]/*him[DAT]
OBJ=ACC
'He finds him.'

b. *Er* *findet* *Papageien.*
he finds parrots[NOM/ACC/DAT/GEN]
OBJ=ACC
'He finds parrots.'

(32) *a.* *Er* *hilft* **ihn/ihm.*
he helps *him[ACC]/him[DAT]
OBJ=DAT
'He helps him.'

b. *Er* *hilft* *Papageien.*
he helps parrots[NOM/ACC/DAT/GEN]
OBJ=DAT
'He helps parrots.'

Groos and Reimsdijk (1979) and Zaenen and Karttunen (1984) were among the first to point out that syncretic forms like *Papageien* can be syntactically **indeterminate** — that is, simultaneously compatible with more than one requirement for a feature such as case.

(33) *Er* *findet* *und* *hilft* *Papageien.*
he finds and helps parrots
OBJ=ACC OBJ=DAT NOM/ACC/DAT/GEN
'He finds and helps parrots.'

In their analysis of indeterminacy, Dalrymple, King, et al. (2009) propose that the value of the CASE attribute is a feature structure which allows specification and differentiation of each (core) case by means of a separate (boolean-valued) attribute: NOM, ACC, DAT, and so forth. A negative value indicates the inability of a form to satisfy the corresponding case requirement, while a positive value indicates that the form can satisfy the requirement. Indeterminate forms can satisfy more case requirements than determinate forms; thus, indeterminate forms contain a smaller number of negative specifications and allow a larger number of positive specifications for case. The value of the CASE feature of the determinately specified German pronouns *ihn* and *ihm* are as given in (34).

(34) Determinate accusative case (*ihn*):

$$\begin{bmatrix} \text{CASE} & \begin{bmatrix} \text{NOM} & - \\ \text{ACC} & + \\ \text{GEN} & - \\ \text{DAT} & - \end{bmatrix} \end{bmatrix}$$

Determinate dative case (*ihm*):

$$\begin{bmatrix} \text{CASE} & \begin{bmatrix} \text{NOM} & - \\ \text{ACC} & - \\ \text{GEN} & - \\ \text{DAT} & + \end{bmatrix} \end{bmatrix}$$

The requirement for the OBJ of *hilft* to bear dative case is imposed by the equation in (35), which requires the value + for the DAT case attribute of *hilft*'s object.

(35) hilft: (↑ OBJ CASE DAT)=+

A dative object like *ihm* can satisfy this case requirement.

(36) *hilft ihm:* DAT

$$\begin{bmatrix} \text{PRED} & \text{'HELP}\langle\text{SUBJ,OBJ}\rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'HIM'} \\ \text{CASE} & \begin{bmatrix} \text{NOM} & - \\ \text{ACC} & - \\ \text{GEN} & - \\ \text{DAT} & + \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

An accusative object like *ihn* fails to satisfy this requirement, since *hilft*'s requirement for DAT + clashes with *ihn*'s requirement for DAT −.

(37) **hilft ihn:* ACC

$$\begin{bmatrix} \text{PRED} & \text{'HELP}\langle\text{SUBJ,OBJ}\rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'HIM'} \\ \text{CASE} & \begin{bmatrix} \text{NOM} & - \\ \text{ACC} & + \\ \text{GEN} & - \\ \text{DAT} & \boxed{+/-} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

An indeterminate form like *Papageien* is a cased form: it must express some case or other, but there are no restrictions on which case it expresses. This means that it can appear as the object of a verb like *findet* (39) as well as a verb like *hilft* (40), since it can be positively specified for either accusative or dative case. As shown in (41), it can also be **overspecified**, with positive values for both features; that is, it can bear more than one case value at the same time.

(38) Papageien: (↑ CASE {NOM|ACC|DAT|GEN})=+

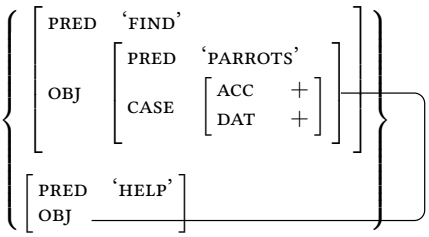
(39) *Er findet Papageien.*
he finds parrots
'He finds parrots.'

$$\begin{bmatrix} \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'PARROTS'} \\ \text{CASE} & [\,\text{ACC} \quad +\,] \end{bmatrix} \end{bmatrix}$$

(40) *Er hilft Papageien.*
he helps parrots
'He helps parrots.'

$$\begin{bmatrix} \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'PARROTS'} \\ \text{CASE} & [\,\text{DAT} \quad +\,] \end{bmatrix} \end{bmatrix}$$

(41) *Er findet und hilft Papageien.*
he finds and helps parrots
'He finds and helps parrots.'



6 Feature-based decomposition of c-structure categories

We assume that the features relevant for c-structure category labels encode only c-structure information: phrase structure category, bar level, functional vs. lexical category, and whether the category is projecting (Toivonen 2003). Bresnan et al. (2015, p. 103) provide a discussion of bar-level features, features distinguishing functional from lexical categories, and features distinguishing projecting and nonprojecting categories; since our aim is to encode indeterminate and determinate constraints on c-structure categories and the category of unlike coordinations, we abstract away from those features, and concentrate only on features that encode phrase structure category.

(42) Nouns and noun phrases are: [N +, V −, P −, ADJ −, ADV −]

Verbs and verb phrases are: [N −, V +, P −, ADJ −, ADV −]

Prepositions and prepositional phrases are: [N −, V −, P +, ADJ −, ADV −]

Adjectives and adjective phrases are: [N −, V −, P −, ADJ +, ADV −]

Adverbs and adverb phrases are: [N −, V −, P −, ADJ −, ADV +]

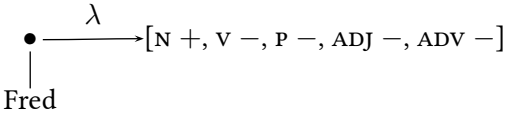
We can now treat N, V, etc. as abbreviations for the corresponding fully instantiated feature matrix:

(43)

Abbreviation	Feature matrix
N	[N +, V −, P −, ADJ −, ADV −]
V	[N −, V +, P −, ADJ −, ADV −]
P	[N −, V −, P +, ADJ −, ADV −]
Adj	[N −, V −, P −, ADJ +, ADV −]
Adv	[N −, V −, P −, ADJ −, ADV +]

And we reinterpret the λ labeling function in (22) as a function from nodes to feature-based node labels.

(44) The node labeling function λ with feature matrices (*Fred* is a noun):



7 Category selection

The following standard notation allows reference to the current node, its mother, and the labels of these nodes.

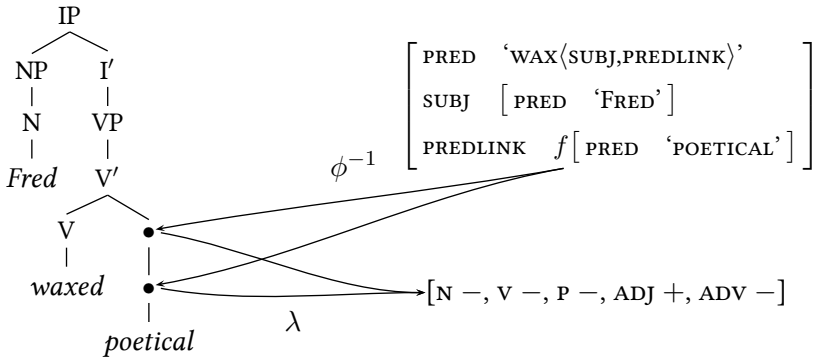
(45) Current node: $*$ Label of current node: $*_\lambda$
Mother node: $\hat{*}$ Label of mother node: $\hat{*}_\lambda$

Lexical entries specify the category of the preterminal node by specifying values for each of the category features. We know of no reason to suppose that the lexicon contains words with overspecified category features, and so we expect all words in the lexicon to have a positive specification for one category feature, and negative specifications for the other category features. Adjectives like *poetical* and *proud* are specified as in (46), with a positive specification for ADJ, and a negative specification for the other category features.

(46) *poetical, proud*
$(\hat{*}_\lambda \text{ N}) = -$
$(\hat{*}_\lambda \text{ V}) = -$
$(\hat{*}_\lambda \text{ P}) = -$
$(\hat{*}_\lambda \text{ ADJ}) = +$
$(\hat{*}_\lambda \text{ ADV}) = -$

On this view, the sentence *Fred waxed poetical* has the following analysis.[3]

(47)



We can now recast our analysis of the requirements of *wax* in feature-based terms: compare (25) with (48). The constraints in (48) use a local variable %C to refer to an arbitrary member of the CAT set of nodes and to specify its required properties. The representation in (47) meets the requirements in (48), as desired.

3 Recall that our focus is on category features, and we ignore features defining bar level. Including a bar level feature would mean that the labels of the two nodes dominating *poetical* would not be the same, since the Adj node would then be distinguished from the AP node by the bar level feature.

(48) Constraints imposed by *wax*:

$\text{CAT}((\uparrow \textsc{predlink}), \%\textsc{c})$

$(\%\textsc{c}\ \textsc{n}) = -$
$(\%\textsc{c}\ \textsc{v}) = -$
$(\%\textsc{c}\ \textsc{p}) = -$
$(\%\textsc{c}\ \textsc{adj}) = +$
$(\%\textsc{c}\ \textsc{adv}) = -$

8 C-structure category of a coordinate phrase

We propose that the category of a coordinate phrase has the value + for a category feature if there is some conjunct with the value + for that feature. On this view, as shown in (49), unlike category coordination involves overspecification: a coordination of unlike categories has the value + for more than one category feature.

(49)

NP:	[N +, V −, P −, ADJ −, ADV −]
AdjP:	[N −, V −, P −, ADJ +, ADV −]
NP and AdjP:	[N +, ADJ +]

Predicates check c-structure category requirements and rule out disallowed options by requiring a negative value for the disallowed feature. For example, an NP or a coordinate phrase containing a NP has the value + for the feature N, and predicates or contexts disallowing NP specify the conflicting value − for the N feature.

9 The coordination rule

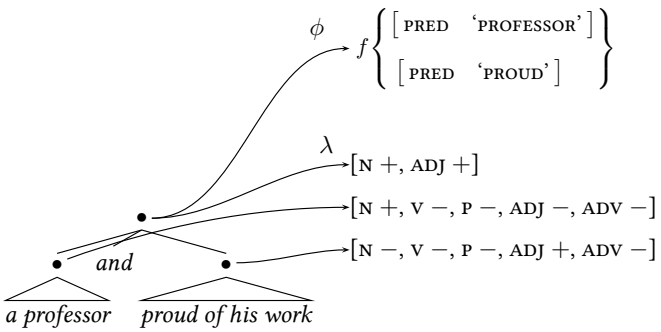
This analysis requires each conjunct daughter to pass up any category feature which has a + value. This is accomplished by annotating each daughter in the coordination rule with the constraints in (50). According to these constraints, if the label (λ) of the daughter node ($*$) has the value + for the feature N, then ($\Rightarrow$) the label of the mother node ($\hat{*}$) is also required to have the value + for the feature N, and similarly for the other category features. If the daughter node has any value other than + for a feature (if it has the value − or is unspecified), nothing is passed up, and the coordinate structure remains unspecified for that feature.

(50) Constraints associated with each daughter node in the coordination rule:

$(*_\lambda\ \textsc{n}) = + \Rightarrow (\hat{*}_\lambda\ \textsc{n}) = +$
$(*_\lambda\ \textsc{v}) = + \Rightarrow (\hat{*}_\lambda\ \textsc{v}) = +$
$(*_\lambda\ \textsc{p}) = + \Rightarrow (\hat{*}_\lambda\ \textsc{p}) = +$
$(*_\lambda\ \textsc{adj}) = + \Rightarrow (\hat{*}_\lambda\ \textsc{adj}) = +$
$(*_\lambda\ \textsc{adv}) = + \Rightarrow (\hat{*}_\lambda\ \textsc{adv}) = +$

These constraints produce the category features [N +, ADJ +] for the unlike category coordination *a professor and proud of his work*, since one of the conjuncts is [N +] and the other is [ADJ +].

(51)



10 Indeterminacy and category selection

10.1 Selection by a predicate

A verb such as *become* places indeterminate requirements on the category of its PREDLINK complement. The category features of the PREDLINK must be compatible with the negative value − for the features V, P, and ADV (they must be unspecified for each of those features, or specified as −), but no constraints on the features N and ADJ are imposed. This means that **either** or **both** of those features can have the value +.

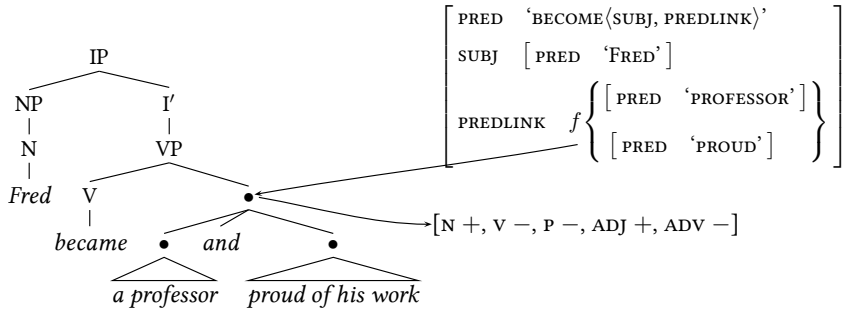
(52) Constraints imposed by *become*:

CAT((↑ PREDLINK), %C)

(%C V) = −
(%C P) = −
(%C ADV) = −

This allows the analysis in (53) of category selection in unlike category coordination. The positive values for the N and ADJ feature come from the coordination rule and the constraints listed in (50), and the CAT constraint in (52) has the effect of negatively instantiating the values of the V, P, and ADV features. In this case, then, the result is that the category of the coordinate structure is fully instantiated, with a value for each category feature.

(53)



10.2 Indeterminate specification in phrase structure rules

As discussed in Section 2.2, a P′ has a head daughter P, and a complement daughter that may be either a nominal phrase or a prepositional phrase. Any of the following are allowed in complement position of a PP:

(54)

NP:	[N +, P −, V −, ADJ −, ADV −]
PP:	[N −, P +, V −, ADJ −, ADV −]
NP and PP:	[N +, P +]

The complement in the P′ rule can have a positive value for the N feature, the P feature, or both, and must be compatible with negative values for the remaining features. The P′ rule can be written as follows, using the abbreviations in (43) for the fully specified categories P′ and P, and an underspecified description for the category of the complement of P.

(55) P′ phrase structure rule, using abbreviations P and P′ for fully instantiated feature matrices:

$$\text{P}' \longrightarrow \text{P} \qquad \begin{array}{c} \bullet \\ (*_\lambda \text{ V}) = - \\ (*_\lambda \text{ ADJ}) = - \\ (*_\lambda \text{ ADV}) = - \end{array}$$

11 Inventory of category features

We have assumed a set of features {N, V, P, ADJ, ADV} that allows maximum differentiation among categories: each feature registers a node as specified for that part of speech. More parsimonious theories of category features have, of course, been proposed; for example, Toivonen (2003) and Bresnan et al. (2015) propose a two-feature system with the features [±PREDICATIVE, ±TRANSITIVE]. The example in (56) is from Bresnan et al. (2015, p. 103).

(56)

	"predicative"	"transitive"
V	+	+
P	−	+
N	−	−
A	+	−

As discussed in detail by Bayer (1996), such decompositions are in general not fine-grained enough to cover all cases of unlike category coordination. In particular, some combinations have no features in common, so it is not possible to use feature underspecification to group together natural classes of all possible combinations. For example, Marcotte (2014) and Bresnan et al. (2015) propose that IP and CP are verbal functional categories, sharing the "predicative" and "transitive" features of verbs but with additional features to mark their status as functional categories. Under their proposal, there is no feature that NPs and CPs have in common.

(57) Pat remembered [the appointment]$_{\text{NP: [PREDICATIVE }-\text{, TRANSITIVE }-]}$ and [that it was important to be on time]$_{\text{CP: [PREDICATIVE }+\text{, TRANSITIVE }+]}$.

The general problem is that the features in (56) are not intended to underpin an analysis of unlike category coordination; the aim is instead to capture a different set of generalizations concerning the relation between functional or lexical categories, or the syntactic combinatory possibilities of the categories (whether they can act predicatively or take an OBJ complement). We use a maximally differentiated feature set in order to be sure that all combinations of categories in unlike category coordinations can be represented and constrained; future work may reveal that a simpler system is possible.

12 Conclusion

This paper addresses one aspect of a general issue that has been the focus of a great deal of attention in the literature. Often, the problem of unlike category coordination is treated as a part of the general problem of syntactic feature resolution and feature indeterminacy, and much of the literature focuses on f-structure features such as case, person, number, and gender; besides the work cited above, relevant work has been done by Pullum and Zwicky (1986), King and Dalrymple (2004), Dalrymple, King, et al. (2006), Dalrymple, Dyvik, and Sadler (2007), and many more. Kaplan (2017) provides an overview discussion of features and underspecification, and proposes a set-based alternative to feature structure-based accounts of indeterminacy which could be explored as an alternative to the account presented here.

In distinguished conjunct agreement, one conjunct in a coordinate structure is syntactically 'distinguished' in that it controls agreement processes (Arnold et al. 2007; Dalrymple and Hristov 2010; Kuhn and Sadler 2007; Sadler 1999, 2003). Sadler (1999)

provides the Welsh examples in (58) to illustrate this pattern: the verb shows first person singular agreement with the first conjunct in example (58a), and third person singular agreement with the first conjunct in example (58b).[4]

(58) *a.* *Roeddwn i a Mair i briodi.*
was.1SG 1SG and Mair to marry
'I and Mair were to marry.'

b. *Roedd Mair a fi i briodi.*
was.3SG Mair and 1SG to marry
'Mair and I were to marry.'

Similar patterns have been claimed to be relevant for c-structure category selection; Sag et al. (1985) discuss examples such as (59), which indicate that the category of the first conjunct can determine the distribution of a coordinate structure (see also Al Khalaf (2015)).

(59) *a.* You can depend on [my assistant] and [that he will be on time].

b. *You can depend on [that he will be on time].

Such examples are actually ruled out by the rule in (55), which constrains all of the categories in a coordinate phrase, and forbids CP arguments in the complement position of a PP; further work is needed to incorporate a treatment of distinguished conjunct category constraints into the overall theory.

Acknowledgments

Helge Dyvik's long-standing interest in the grammar of features in coordination is the inspiration for the proposal presented here, which also builds on our earlier work with Louisa Sadler on gender resolution in coordination (Dalrymple, Dyvik, and Sadler 2007), and I am very happy to dedicate this paper to him. I am also grateful to Miriam Butt, Jamie Findlay, Ron Kaplan, Tracy King, Stephen Jones, Joey Lovestrand, John Lowe, John Maxwell, Adam Przepiórkowski, Louisa Sadler, and two anonymous reviewers for helpful comments.

References

Al Khalaf, Eman (2015). "Coordination and Linear Order". PhD thesis. University of Delaware.

Arnold, Doug, Louisa Sadler, and Aline Villavicencio (2007). "Portuguese: Corpora, Coordination and Agreement". In: *Roots: Linguistics in Search of its Evidential Base*. Ed. by Sam Featherston and Wolfgang Sternefeld. Berlin: Mouton de Gruyter, pp. 9–28.

4 Thanks to Louisa Sadler for discussion of this point.

Bayer, Samuel (1996). "The Coordination of Unlike Categories". In: *Language* 72.3, pp. 579–616.

Beavers, John and Ivan A. Sag (2004). "Coordinate Ellipsis and Apparent Non-Constituent Coordination". In: *Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar*. Ed. by Stefan Müller. Stanford: CSLI Publications, pp. 48–69.

Bresnan, Joan (1974). "The Position of Certain Clause-Particles in Phrase Structure". In: *Linguistic Inquiry* 5.4, pp. 614–619.

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler (2015). *Lexical-Functional Syntax*. Oxford: Wiley Blackwell.

Butt, Miriam, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond (1999). *A Grammar Writer's Cookbook*. Stanford: CSLI Publications.

Crouch, Dick, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula S. Newman (2008). *XLE Documentation*. Palo Alto Research Center. Palo Alto, CA.

Dalrymple, Mary, Helge Dyvik, and Tracy Holloway King (2004). "Copular Complements: Closed or Open?" In: *On-Line Proceedings of the LFG2004 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Dalrymple, Mary, Helge Dyvik, and Louisa Sadler (2007). "Gender Resolution in Coordination and Disjunction". Abstract in *On-Line Proceedings of the LFG2007 Conference*, ed. Miriam Butt and Tracy Holloway King.

Dalrymple, Mary and Bozhil Hristov (2010). "Agreement Patterns and Coordination in Lexical Functional Grammar". In: *On-Line Proceedings of the LFG2010 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Dalrymple, Mary and Ronald M. Kaplan (2000). "Feature Indeterminacy and Feature Resolution". In: *Language* 76.4, pp. 759–798.

Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, eds. (1995). *Formal Issues in Lexical-Functional Grammar*. Stanford: CSLI Publications.

Dalrymple, Mary, Tracy Holloway King, and Louisa Sadler (2006). "Indeterminacy by Underspecification". In: *On-Line Proceedings of the LFG2006 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

– (2009). "Indeterminacy by Underspecification". In: *Journal of Linguistics* 45, pp. 31–68.

Dalrymple, Mary and Helge Lødrup (2000). "The Grammatical Functions of Complement Clauses". In: *On-Line Proceedings of the LFG2000 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Davies, Mark (2013). *Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day*.

– (2015). *The Wikipedia Corpus: 4.6 million articles, 1.9 billion words*. Adapted from Wikipedia.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, and Ivan A. Sag (1985). *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard University Press.

Groos, Anneke and Henk van Reimsdijk (1979). "Matching Effects in Free Relatives: A Parameter of Core Grammar". In: *Theory of Markedness in Generative Grammar: Proceedings of the 1979 GLOW Conference*. Ed. by Adriana Belletti, Luciana Brandi, and Luigi Rizzi. GLOW. Pisa: Scuola Normale Superiore di Pisa, pp. 171–216.

Jacobson, Pauline (1987). "Review of G. Gazdar, E. Klein, G. Pullum, and I. Sag, *Generalized Phrase Structure Grammar*". In: *Linguistics and Philosophy* 10.3, pp. 389–426.

Kaplan, Ronald M. (2017). "Formal Aspects of Underspecified Features". In: *Festschrift for Lauri Karttunen*. Ed. by Cleo Condoravdi and Tracy Holloway King. Stanford: CSLI Publications.

Kaplan, Ronald M. and Joan Bresnan (1982). "Lexical-Functional Grammar: A Formal System for Grammatical Representation". In: *The Mental Representation of Grammatical Relations*. Ed. by Joan Bresnan. Reprinted in Dalrymple, Kaplan, Maxwell, and Zaenen (1995, pp. 29–130). Cambridge, MA: The MIT Press, pp. 173–281.

Kaplan, Ronald M. and John T. Maxwell III (1996). *LFG Grammar Writer's Workbench*. Tech. rep. Palo Alto, CA: Xerox Palo Alto Research Center.

Kaplan, Ronald M. and Annie Zaenen (1989). "Long-Distance Dependencies, Constituent Structure, and Functional Uncertainty". In: *Alternative Conceptions of Phrase Structure*. Ed. by Mark R. Baltin and Anthony S. Kroch. Reprinted in Dalrymple, Kaplan, Maxwell, and Zaenen (1995, pp. 137–165). Chicago: University of Chicago Press, pp. 17–42.

King, Tracy Holloway and Mary Dalrymple (2004). "Determiner Agreement and Noun Conjunction". In: *Journal of Linguistics* 40.1, pp. 69–104.

Kuhn, Jonas and Louisa Sadler (2007). "Single Conjunct Agreement and the Formal Treatment of Coordination in LFG". In: *On-Line Proceedings of the LFG2007 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Marcotte, Jean-Philippe (2014). "Syntactic Categories in the Correspondence Architecture". In: *On-Line Proceedings of the LFG2014 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Patejuk, Agnieszka (2015). "Unlike Coordination in Polish: An LFG Account". PhD thesis. Polish Academy of Sciences.

Peterson, Peter G. (2004). "Coordination: Consequences of a Lexical-Functional Account". In: *Natural Language and Linguistic Theory* 22.3, pp. 643–679.

Pollard, Carl and Ivan A. Sag (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Pullum, Geoffrey K. and Arnold M. Zwicky (1986). "Phonological Resolution of Syntactic Feature Conflict". In: *Language* 62.4, pp. 751–773.

Sadler, Louisa (1999). "Non-Distributive Features and Coordination in Welsh". In: *On-Line Proceedings of the LFG99 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

– (2003). "Coordination and Asymmetric Agreement in Welsh". In: *Nominals: Inside and Out*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications, pp. 85–118.

Sag, Ivan A. (2002). "Coordination and Underspecification". In: *Proceedings of the 9th International Conference on Head-Driven Phrase Structure Grammar*. Ed. by Jong-Bok Kim and Stephen Wechsler. Stanford: CSLI Publications, pp. 267–291.

Sag, Ivan A., Gerald Gazdar, Thomas Wasow, and Steven Weisler (1985). "Coordination and How to Distinguish Categories". In: *Natural Language and Linguistic Theory* 3.2, pp. 117–171.

Toivonen, Ida (2003). *Non-Projecting Words: A Case Study of Swedish Verbal Particles*. Dordrecht: Kluwer.

Zaenen, Annie and Lauri Karttunen (1984). "Morphological Non-Distinctiveness and Coordination". In: *Proceedings of the Eastern States Conference on Linguistics (ESCOL '84)*. Ed. by Mark Cobler, Susannah MacKaye, and Michael T. Wescoat, pp. 309–320.

Finite-state tokenization for a deep Wolof LFG grammar

Cheikh M. Bamba Dione

Abstract. This paper presents a finite-state transducer (FST) for tokenizing and normalizing natural texts that are input to a large-scale LFG grammar for Wolof. In the early stage of grammar development, a language-independent tokenizer was used to split the input stream into a unique sequence of tokens. This simple transducer took into account general character classes, without using any language-specific information. However, at a later stage of grammar development, uncovered and non-trivial tokenization issues arose, including issues related to multi-word expressions (MWEs), clitics and text normalization. As a consequence, the tokenizer was extended by integrating FST components. This extension was crucial for scaling the hand-written grammar to free text and for enhancing the performance of the parser.

1 Introduction

This paper presents a finite-state transducer (FST) (Beesley and Karttunen 2003) that acts as a tokenizer and a normalizer for Wolof[1] natural texts. Tokenization constitutes an important prior task for various language processing applications, e.g. part-of-speech tagging, parsing, information retrieval, information extraction, and machine translation. All these language processing systems need input texts with definite word boundaries. This task can be performed using various techniques, including e.g. rule-based techniques (Kaplan 2005), statistical techniques (Yang and Li 2005) and lexical techniques (Wu and Fung 1994). The tokenization approach proposed in this paper is based on the use of finite-state rules to break up a stream of Wolof texts into individual tokens. The tokenizer is designed using the Xerox finite-state tool *fst* (Beesley and Karttunen 2003).

The tokenization system is built within the broader context of an ongoing process of creating language resources and tools for Wolof. This process is part of the Parallel Grammar (ParGram) project (Butt et al. 2002) which is couched within the Lexical Functional Grammar (LFG) framework. In related work, a Wolof Morphological Analyzer (WoMA) (Dione 2012), a large-scale LFG grammar and a treebank for Wolof

1 Wolof belongs to the Senegambian branch of the Niger–Congo language family mainly spoken in Senegal, Gambia and Mauritania.

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1340.

have been constructed (Dione 2014a). Other related work describes parse disambiguation techniques used for Wolof (Dione 2014b), including the integration of Constraint Grammar (CG) models (Karlsson 1990) into probabilistic context-free grammar approaches to disambiguation (Dione 2014c).

The tokenizer is used as part of a finite-state transducer cascade (Kaplan et al. 2004) that preprocesses the input sentences before they are parsed by the Wolof grammar.

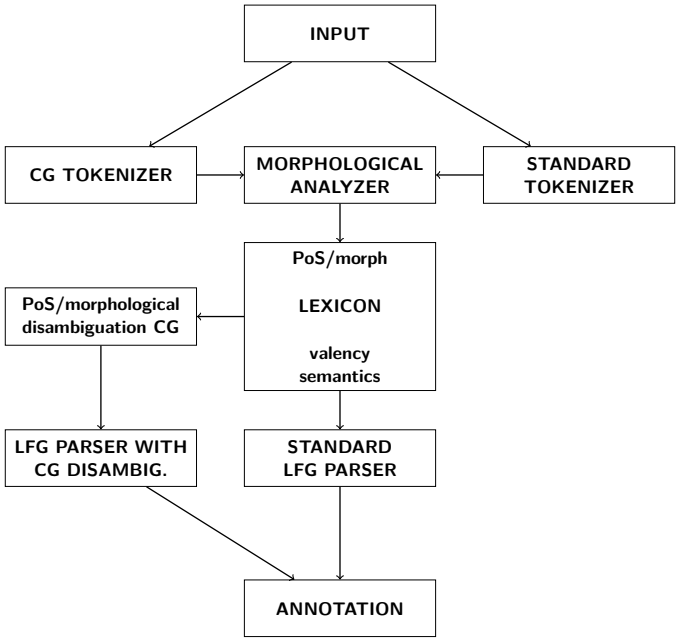


Figure 1: Anatomy of the Wolof parsing system

The parsing workflow is depicted in Figure 1. First, the input is tokenized and normalized either by a deterministic tokenizer when the LFG parsing is combined with CG disambiguation (Dione 2014c), or by a deterministic tokenizer when the syntactic analysis is performed without CG disambiguation. The former tokenizer is referred to as the CG tokenizer, while the latter is called the standard tokenizer. The only difference between the two tokenizers is their determinism. Next, morphological analysis is carried out. The output of the morphology is either disambiguated prior to syntactic analysis or directly fed into the standard LFG parser (i.e. without CG disambiguation). Finally, the morpho-syntactic annotation is produced.

In the early stage of grammar development, a language-independent tokenizer was used by the Wolof LFG system. However, as the development of the grammar progressed, the parsing system encountered various issues due to inappropriate tokenization. For instance, a significant number of sentences with MWEs could not be parsed or were not handled correctly. Likewise, the time needed to process sentences including

words with clitics was growing due to parsing complexity. Hence, the need to integrate language-specific information into the tokenization process arose naturally with the aim to enhance coverage and quality as well as efficiency of the parser.

The remainder of this paper is organized as follows: Section 2 discusses the general concept of tokens in Wolof and non-trivial tokenization issues found in this language. Section 3 discusses the development of the tokenizer using finite-state technology. It presents the language-independent tokenizer used at the early stage of grammar development and describes the final transducer which integrates language-specific approaches to multi-word expressions and clitics into the tokenization process. Section 4 discusses issues related to text normalization. Section 5 reports on results of experimental evaluation of the tokenizer. Section 6 concludes the discussion.

2 Wolof tokens

Tokenization can be defined as the process of breaking a stream of texts up into words, symbols, or other meaningful elements called tokens. Accordingly, the process is assumed to typically occur at the word / token level. However, in some cases, it may be difficult to exactly define what is meant by a 'word' or 'token'. This is particularly true for an agglutinative language like Wolof.

Similar to Turkish (Oflazer et al. 2004), many derivational phenomena in Wolof take place within a word form, but there are other complex derivations involving compounds and reduplications (Ka 1994). Wolof word forms consist of morphemes concatenated to a root morpheme or to other morphemes. The language is almost exclusively suffixing. In many contexts, the surface realizations of the morphemes are conditioned by various morphophonemic processes such as vowel harmony, vowel and consonant elisions, gemination, degemination, vowel coalescence, glide insertion, prenasalization, etc. (Ka 1994).

The morphotactics of Wolof word forms can be quite complex when multiple derivations and inflections are involved. For instance, the verb *gënoonatee* in (1) which consists of different derivational and inflectional morphemes can be represented as in (2).

(1) *Li* *gën-oon-ati-a* *metti*
What SURPASS-PST-Iter-Cinf be.painful
'Particularly painful was … '

(2) gënoonatee ⇔ gën+Verb+Modal+Past+Iter+A

This word starts out with a root *gën* 'be better/worse' followed by the past tense marker *-oon*, the iterative suffix *-ati* and the infinitival complementizer *a* which surfaces as a clitic. The ending *-atee* results from a vowel coalescence process: the final vowel of the suffix *-ati* is collapsed with the clitic *a*. Without derivation and inflection, the contraction of the verb in (2) with the the infinitival complementizer (Cinf) *a* can be

tokenized at least in two different ways: either by handling the clitic as a normal affix integrated into the verbal stem (i.e. *gëna*) or by demarcating it from the stem (i.e. *gën a*). Accordingly, a precise definition of the *token* concept in Wolof is required, before an accurate tokenization system can be built for this language.

Throughout this research work, the definition of a token follows the one given for Arabic by Attia (2007, p. 66): "the minimal syntactic unit; it can be a word, a part of a word (or a clitic), a multiword expression, or a punctuation mark". Accordingly, two categories of tokens can be distinguished for Wolof: *main tokens* vs. *sub-tokens*. *Main tokens* refer to stems with or without clitics, as well as numbers, which are typically separated by white spaces and punctuation marks as delimiters or word boundaries. Also, single character symbols like quotation marks and punctuation used in Wolof, such as the period, comma, question mark, semicolon, etc., are treated as individual tokens. In contrast, in some other cases (see Table 1), a stem may be suffixed with a clitic, both represented as *sub-tokens*.

2.1 Wolof clitics

As discussed in the previous section, a challenging tokenization issue is cliticization. Like Arabic (Attia 2007), Wolof morphotactics allows words to be suffixed with clitics. Clitics themselves can be concatenated one after the other. Furthermore, clitics undergo assimilation with word stems and with each other, making it difficult to recognize and handle them properly. Examples of full form words consisting of stems with clitics are shown in Table 1. Assimilation can be observed in some of these examples. The first row of the table is to be read as follows: the preposition *ak* 'with' may encliticize to the verbal stem *daje* 'meet', yielding the surface form *dajeek*.[2] The other surface forms involve different grammatical categories (determiners, conjunctions, pronouns, etc.) and occur in a similar manner.

2.2 The use of multiword expressions

Another relevant tokenization issue is the use of multiword expressions (MWEs). Formally, MWEs can be defined as "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al. 2002, p. 90). More specifically, MWEs are "two or more words that behave like a single word syntactically and semantically" (Attia 2007, p. 68). MWEs can be of different types, including idioms, prepositional verbs, verbs with particles, collocations, etc. Following Attia (2007), Oflazer et al. (2004), and Sag et al. (2002), Wolof MWEs are classified into four types: named entities, fixed expressions, semi-fixed expressions and syntactically flexible expressions.

1. Multi-word named entities refer to proper nouns for persons, organizations, places, etc., as illustrated in (3).

2 The long vowel *ee* in the surface form *dajeek* results from a vowel coalescence: the final vowel of the verbal stem *-e* coalesces with the stem-initial vowel of the preposition, i.e. *-a*.

Stem	*Clitic category*	*Example*	*Word form*	*Literal translation*
Verb	PREP	*daje* 'meat' + *ak* 'with'	*dajeek*	'met with'
	DET	*joxe* 'give' + ay 'some'	*joxeey*	'give some'
	Inf. COMP	*soog* 'start' + *a*	*sooga*	'start to V'
Determiner	PREP	*ba* 'the' + *ak* 'with'	*baak*	'the with'
	CONJ	*bi* 'the' + *ak* 'and'	*beek*	'the and'
Preposition	DET	*ci* 'in' + *ab* 'a'	*cib*	'in a'
	PREP	*ca* 'about' + *ak* 'with'	*caak*	'about with'
Noun	CONJ	*ndox* 'water' + *ak* 'and'	*ndoxak*	'water and'
Name	CONJ	*Ali* 'Ali' + *ak* 'and'	*Aleek*	'Ali and ...'
Adverb	PRON	*fu* 'where' + *nga* 'you'	*foo*	'where you ...'
Complementizer	PRON	*bu* 'if' + *nga* 'you'	*boo*	'if you ...'
Pronoun	CONJ	*moom* 'him' + *ak* 'and'	*mook*	'... and him'
Object pronoun	Inf. COMP	*ko* 'him/her' + *a*	*koo*	'him/her' + inf. V
Conjunction	AUX	*te* 'and' + *di* imperf.	*tey*	'and' + imperf.
	DET	*mbaa* 'or' + *ay* 'some'	*mbaay*	'or some'

Table 1: Examples of Wolof stems from different grammatical categories with clitic sub-tokens

(3) *Daara ju Kowe ji*
school REL high DET
'The University' (Lit. 'The school which is high')

2. Fixed expressions denote *collocations* where all components of the collocation are lexically, syntactically and morphologically rigid, as in (4). Other Wolof fixed MWEs include adverbials *saa su ne* 'every time', and quantifying expressions *bu baax* 'very well', etc. None of these MWEs can be reordered or separated by external elements.

(4) Mag ak rakk 'siblings'

3. Semi-fixed expressions refer to collocations where some components of the collocation are fixed and some can vary. The variation can be of morphological (e.g. inflectional or derivational) or lexical type (where one word can be replaced by another). Wolof inflectional subject markers like *maa ngi* in (5) and *noo ngi* in (6) are instances of semi-fixed expressions. They vary according to person and number of the subject as well as the aspect of the verb, as illustrated in (5)–(6), the optional attachment of the imperfective marker *-y* indicates the collocation

being inflected for aspect. In contrast, (6) exemplifies a case of lexical variation, where the word *maa* in (5) is replaced by *noo*.

(5) *Maa ngi(-y) dem*
1sg.PROG-IPF go
'I am leaving (here and now)'

(6) *Noo ngi dem*
1pl.PROG go
'We are leaving'

4. Syntactically flexible expressions are non-lexicalized expressions that can undergo reordering or allow external elements to intervene between the components of the collocation. For example, the MWE in (7) is disrupted in (8) by the agreement marker *na* and the clitic pronoun *ko*.

   (7) fas yéene 'decide'

   (8) fas na ko yéene 'He has decided it'

Wolof multiword tokens can be of different grammatical categories: inflectional elements (5); adverbial expressions like *bu baax* 'very well' (10) and *saa su ne* 'every time' (9); prepositions like *ci biir* 'inside'; pronouns such as *yoo xam ne* 'that/which'; nouns such as *mag ak rakk* 'brothers'; quantifiers like *ku ne* 'every one'; reduplicated words like *jékki jékki* 'suddenly', and other units. Some Wolof examples including multiword expressions are given in (9) and (10).

(9) ***Saa su ne***, ***noo ngi*** *nekk* ***ci biir*** *kër gi.*
Every time 1pl.PROG be inside house the
'Every time, we are inside the house.'

(10) ***Ku ne*** ***jékki jékki*** *xàqtaay* ***bu baax***.
Every one suddenly laugh.out very well.
'Suddenly, every one laughs out loud.'

3 The Wolof tokenizer

This section describes the development of the Wolof tokenizer in finite-state technology. Section 3.1 presents the language-independent tokenizer used in the early stage of grammar development. Section 3.2 discusses the integration of language-specific information into the tokenizer.

3.1 Language-independent tokenization

As noted above, in the early stage of grammar development, tokenization was carried out by a language-independent FST. Typically, a language-independent tokenizer

is a simple kind of deterministic tokenizer, i.e. an unambiguous finite-state transducer which relies on simple heuristics and takes into account some general character classes. For instance, it assumes that contiguous strings of alphabetic characters are part of one token; likewise with numbers. Tokens are separated by whitespace characters (designated by the category `WS`), including e.g. space (`SP`) or line break (`NL`), or by punctuation marks. Accordingly, the `Token` transducer (11) was defined as the union of sequences of alphabetic characters (`WORD`), numbers (`NUM`), punctuation and some other symbols (`SYMB`).

(11)
```
define Token [WORD | SYMB | NUM];
```

In addition, a language-independent tokenizer generally needs to normalize white space. This is because, in natural texts, the use of white space may be uneven and sometimes very inconsistent. For instance, one may find two or more white-space characters, including space, tab, newline characters, etc., instead of a single space. Similarly, spaces might inadvertently be added before or after punctuation marks. Therefore, normalizing tools for eliminating such inconsistencies are needed at a preliminary stage of tokenization. Normalizing the input before processing it allows for the separation of concerns, because the input is assumed to be consistent before operations are performed on it.

With the token definition (11), a language-independent tokenizer that inserts newlines to mark token boundaries (`TB`) can be compiled from the regular expression in (12). It represents the composition of three simple replace terms.

(12)
```
WS+ @-> SP
.o. Token @-> ... NL
.o. [WS]+ & $[NL] @-> TB
```

The first term in (12) reduces strings of whitespace characters into a single blank using longest-match replacement. The second term inserts a newline as a token boundary after the longest matches of letter sequences and other non-whitespace sequences. The third term denotes a rule that removes a set of spaces by replacing it with a token boundary. This set represents the intersection or conjunction of one or more spaces and the set of strings that contains at least one instance of newline somewhere.

When the white space normalizer is fed an input like (13), in which additional spaces are inserted and some spaces are misplaced, it corrects the errors and gives the output in (14).

(13)
Xale	*yi*	*lekk*	*jën*	*wi.*
child	the.pl	eat	fish	the.sg

'So, the children eat the fish.'

(14) Xale yi lekk jën wi.

When surface strings like (14) are looked up using this model, the output string is the input string plus the multi-character symbols TB, inserted between tokens as in (15).

(15) Xale yi lekk jën wi. ⇒ Xale TB yi TB lekk TB jën TB wi TB . TB

However, this language-independent tokenizer encountered serious problems with respect to contracted words, hyphenated words and multiword expressions, as illustrated by examples (16) and (17). For instance, a MWE like *Saa su ne* in (9) was tokenized as shown in (16). However, in an appropriate tokenization model, it should be tokenized as in (17), neglecting the space between the individual tokens when assigning token boundaries. Conversely, an appropriate model would identify the vowel coalescence process involved in (18) and demarcate the collapsed vowels by inserting a space between them, as shown in (19).

(16) Saa su ne ⇒ Saa TB su TB ne TB

(17) Saa su ne ⇒ Saa su ne TB

(18) gënoonatee ⇒ gënoonatee TB

(19) gënoonatee ⇒ gënoonati TB a TB

Accordingly, more sophisticated tokenization techniques were needed to account for these non-trivial issues.

3.2 Integrating language-specific information into the tokenization process

As a result of the tokenization problems in using a language-independent model, I decided to integrate FST components for handling MWEs and clitics into the tokenization system. Similarly, preprocessing tools for text normalization were added. The final architecture of the Wolof tokenization system is depicted in Figure 2.

As Figure 2 shows, the internal preprocessing workflow consists of a cascade of transducers. Thus, during tokenization, the input is first normalized. Then string-based multi-word identification for named entities, fixed expressions and semi-fixed expressions is performed, allowing morphological variation for the latter MWE group. Next, the input is normalized again in order to remove space and to lowercase the first word in a sentence. After, clitics detection and demarcation are performed either deterministically by a clitic transducer, or indeterministically by a clitic guesser. Finally the tokenized and normalized output is produced in two variants according to whether it will be fed into the standard parsing system or the one based on CG disambiguation.

Note that the clitic transducer proposes analyses for contracted words using very basic morphological information carried out by an internal component of the tokenizer. For the standard parsing system, this step is non-deterministic and carried out by a

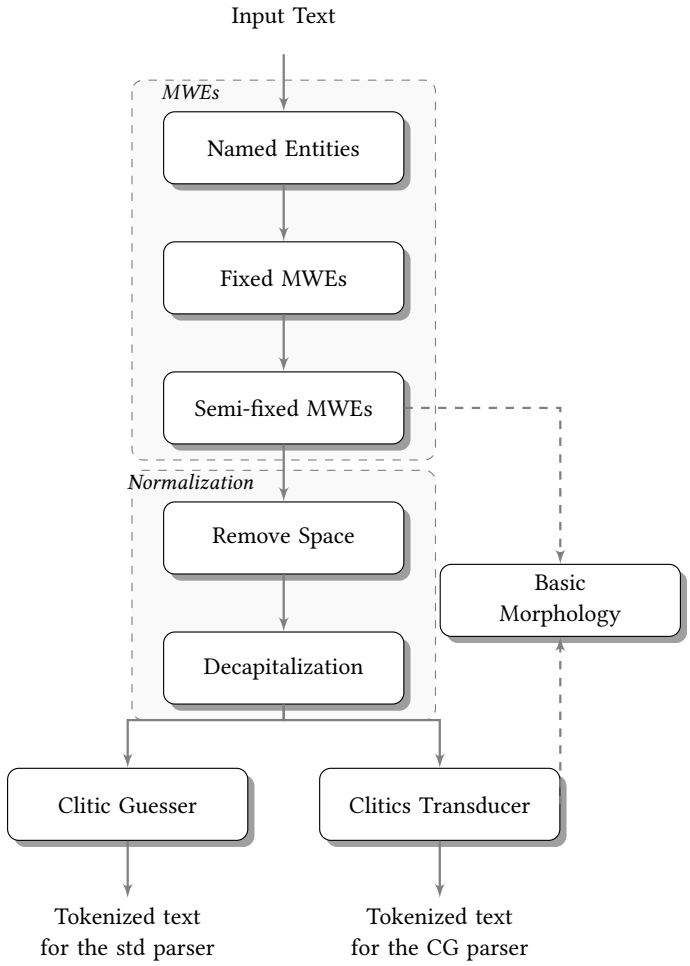


Figure 2: Architecture of the Wolof tokenization system

clitic guesser, since the goal is to allow all possible tokenizations as candidates for syntactic analysis. However, due to the nondeterministic nature of a guesser, there will be increased tokenization ambiguities. In contrast, when parsing is combined with CG disambiguation, this step is deterministic. Therefore, the clitics are not guessed, but rather handled by a transducer carefully designed to produce unambiguous outputs. The individual tokenizing and normalizing components are described in the next sections.

3.2.1 Multiword transducer

Parsing MWEs requires "a deep analysis that starts as early as the normalization and tokenization, and goes through morphological analysis and into syntactic rules" (Attia 2008, p. 70). Handling MWEs at the early preprocessing stage presents some advantages in that it avoids needless analysis of idiosyncratic structures. Additionally, it allows a reduction of ambiguity and parsing time. A precise treatment of MWEs is, however, challenging in that it requires adequate strategies (Beesley and Karttunen 2003). For instance, with a naive model, multi-word tokens may be recognized even when they are just part of a longer alphabetic string, leading to inappropriate tokenization. Therefore, the model used in the present work has been designed such that it will handle them as accurately as possible.

Using regular expressions, a two-sided transducer was created for handling the three following types of MWEs: named entities, fixed expressions and semi-fixed expressions.[3] This transducer was then embedded in the tokenizer as described by Beesley and Karttunen (2003).

In the first stage, finite-state networks including named entities (proper names, locations, organizations, etc.), fixed expressions and semi-fixed expressions were created. The networks represent lists of words separated by space. The lists were created according to the grammatical categories of the MWEs.[4] For instance, for each part of speech such as nouns, adverbs, prepositions, etc., a corresponding finite-state network was built. Unlike named entities and fixed MWEs, the handling of semi-fixed expressions needs some very basic morphological information (see Figure 2) due to morphological variations. Such information was explicitly encoded in the tokenizer as an FST that generates the possible inflectional forms for these MWEs before the final compilation.

In the second stage, a main MWE was built by concatenating all the different finite-state networks created so far. In the third stage, special brackets (e.g. `M1` and `M2`) were inserted around maximally long multi-word expressions, as shown in (20). Subsequently, the main MWE transducer `MWE1` was integrated into the tokenizer, as illustrated in (21). With this rule, the initial *token* concept given in (11) was redefined and augmented with information about MWEs.

(20) `define MWE1 [M1 MWE M2];`

(21) `define Token [WORD | SYMB | NUM | MWE1];`

Finally, the rule in (22) was used to identify multi-word tokens on the basis of information provided by the previous terms. This rule represents a composition of rules. The

3 Note that the MWE transducer was not responsible for the treatment of syntactically flexible expressions, which are handled by the Wolof grammar. As noted above, such expressions are not included in the tokenizer, since their structures allow external (e.g. pronominal) elements to intervene. For more details on how such verbs with particles are handled see e.g. Dione (2014b).

4 The lists included in the MWE transducer are of a moderate size and mostly include multi-word tokens found in the corpus.

first part in (22) considers MWE as those tokens delimited (bounded) by the elements defined in (24). These consist of single character symbols `SINGLE` as defined in (23), or whitespace or the boundary symbol.[5] The second part in (22) then inserts a newline as a token boundary after longest matches of letter sequences and other non-whitespace sequences.

(22)
```
define TOK1 [ MWE @-> M1 ... M2
                || Bound _ Bound
                .o. Token @-> ... NL
                .o. [M1|M2] -> 0
            ];
```

(23)
```
define SINGLE [%"|%«|%»|%.|%,|%;|%:|%!|%?|%(|%)|%[|%]|%{|%}|%–];
```

(24)
```
define Bound [SINGLE|WS|.#.];
```

The third part in (22) removes the special brackets around multiword expressions when they are no longer used. However, as Beesley and Karttunen (2003) pointed out, the brackets also need to be deleted from the sigma alphabet[6] in order to make the input side of the resulting transducer match the universal language. This is accomplished by (25), which explicitly 'absorbs' the auxiliary symbols into the unknown alphabet by calling the Xerox *substitute* commands for each of the brackets used.

(25)
```
substitute symbol ? for "<<"
substitute symbol ? for ">>"
```

With this extended tokenization model, a sentence like (26) will have the phrase structure shown in Figure 3.

(26) *mu dem ci Daara ju Kowe ji*
3sg go PREP university
'He went to the university'

3.2.2 Clitic transducer

In the extended model, tokenization is also expected to handle clitic boundaries. For this purpose, a clitic transducer is embedded into the tokenizer, in a similar manner to the MWE transducer. The clitic transducer has the functions of detecting and demarcating contracted morphemes and handling them as separate words. Yet information on what may constitute a clitic is still needed. One possibility is to use a clitic guesser, making assumptions about clitic occurrence and position in Wolof word formation.

5 The boundary symbol `.#.` denotes the beginning of a string in the left context and the end of the string in the right context of a replace expression.

6 The sigma alphabet is the set of individual symbols known to the network.

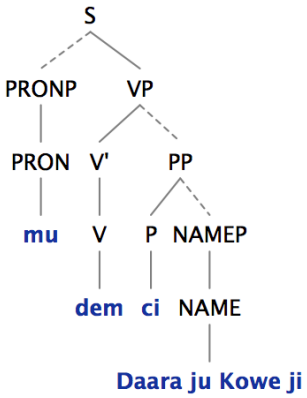


Figure 3: Phrase structure of (26)

For instance, joined words like *baak* (see Table 1) are decomposed into *ba TB ak TB*. The long vowel *aa* is produced by a vowel coalescence rule which collapses the final vowel of the determiner stem *b-* and the initial short vowel of the conjunction *ak* 'and'. This kind of contraction is very common for Wolof determiners, demonstratives, pronouns, etc., which take the noun class index (e.g. *b, g, j, k, l, m, s, w*, etc.).[7] Thus, abstracting from the noun class index, one can formulate a non-deterministic rule like (27) which optionally inserts a token boundary between the collapsed vowels if the morpheme *aak* is found at the end of a word.

(27) `{aak} (->) [a] TB [a k] || _ [.#.|TB]`

Such an operation may be particularly useful when applied to constituents like (28), with the coordinated nouns tokenized as given in Figure (4).

(28) *Petu* *ma-ak* *yedd* *ya-ak* *xuloo* *ba-ak* *lépp*
meeting det-conj lecture det-conj dispute det-conj all
'The secret meetings, the lectures, the disputes and all this'
Lit.: 'The secret meetings and the lectures and the disputes and all this'

This solution, however, is based on a guessing mechanism which naturally increases ambiguity due to non-determinism. Furthermore, since the CG disambiguator needs unambiguous input, such rules cannot be included in the CG tokenizer. Also, because

7 Wolof is a noun class language with noun class agreement (McLaughlin 2010). The language has approximately 13 noun classes identified by their index: 8 singular (*b, g, j, k, l, m, s, w*), 2 plural (*y, ñ*), 2 locative (*f, c*), 1 manner (*n*).

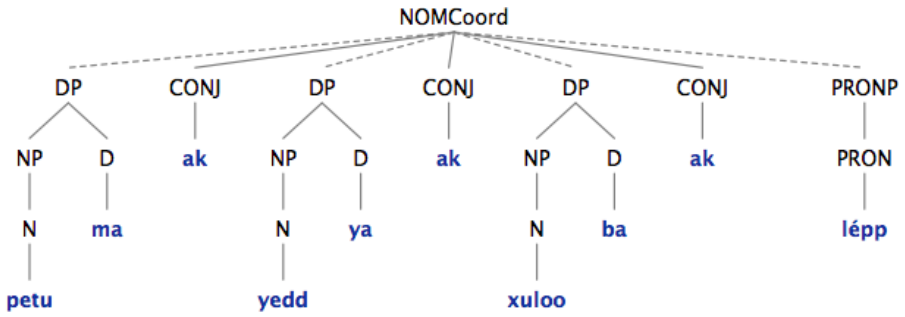


Figure 4: Space insertion using the clitic transducer

cliticization can occur after several morphophonological processes, a proper treatment of clitics needs, for example, morphological information related to verbal inflection and all possible assimilations.

For this reason, a clitic transducer is integrated into the tokenizer. As illustrated in Figure 2, the transducer is associated with an internal FST component that provides basic morphophonological information, e.g. about the forms involved in word contractions as well as the contexts where these occur. For instance, the simplified finite-state network in (29) contains the word stems of modal verbs and other verb operators in Wolof (Church 1981). These stems are not guessed, but taken as a list of actual words extracted from the morphological lexicon.

(29) `define ModalStems [ {soog} | {mën} | {mas} | {bëgg} | {gën} | ... ];`

Such information can be encoded in a finite-state transducer like (30), which represents the composition of the finite-state network `Inflection` with the transducer for vowel coalescence `vowCoal` (both not displayed here).

(30) `define InflModal [ModalStems Inflection] .o. vowCoal;`

Using this information, a clitic transducer like (31) can then detect clitics and insert space between the sub-tokens.

(31)
```
define splitA a -> TB [a] || [.#.|TB]
                      [ModalStems | InflectedModal] _ [.#.|TB]
                      .o.   ...;
```

In (31), a token boundary is unambiguously inserted between the stem of an optionally inflected modal verb and the complementizer clitic *a* found at the end of the verbal string, as illustrated in (32) and (33).

(32) gëna ⇒ gën TB a TB

(33) gënoonatee ⇒ gënoonati TB a TB

4 Text normalization

Besides the integration of language-specific information into the tokenizing transducer, the Wolof FST also includes text and word normalization components. Beesley and Karttunen (2003, p. 440) define word normalization as "the general process of mapping accidental spelling variations to yield normalized forms for analysis". Most commonly needed normalizations in natural-language processing are those that handle initial capitalization (upper-casing) and whole-word capitalization. In Wolof, decapitalization at the beginning of a sentence has proven to be an important issue, as discussed in the next section.

4.1 The initial-capitalization normalizer

Besides the normalization of whitespace shown in section 3.1, decapitalization is a relevant normalization issue for grammar engineering. As Forst and Kaplan (2006, p. 370) noted, "the most important normalization when parsing free text is decapitalization at the beginning of a sentence, **but** also after opening quotes, brackets, colons and hyphens". Accordingly, the tokenizer includes a component for lower-casing accidental spellings, which reflects the orthographical convention of capitalizing the first letter of the first word in a sentence. This kind of normalization is carried out at two different levels: the first word of the sentence is marked by the tokenizer and then lowercased by the morphological analyzer. Note, however, that this normalization form does not apply to proper names, which are considered as a special word category. Proper names are entered in the lexicon with initial capital letter and this spelling will be preserved.

4.1.1 Normalization dependent on the tokenizer

At the tokenization level, the transducer in (34) is used to mark the first word of the sentence. Such a word begins with a capital letter *upper* defined as a network that consists of all the capital letters in Wolof. This word is expected to be found at the beginning of a string (cf. the boundary symbol `.#.`) or after colons followed by one or more token boundaries and optionally symbols occurring before the first word of a sentence (*beforeFirstWord*), e.g. quotes, dashes, parentheses, etc. Accordingly, the transducer inserts the caret symbol before that word.[8] In this particular case, this symbol is used as a hint for the morphological analyzer to identify the word marked as the first one of a sentence.

8 The notation with an initial caret (or hat or circumflex) symbol ^ follows the convention of encoding feature-like multi-character symbols in Xerox finite-state systems (Beesley and Karttunen 2003).

```
(34)  define mark1stWord [ "\^{}" ... <- upper
                           [.#. | ":" TB* ]
                           beforeFirstWord* TB* _
                         ];
```

4.1.2 Normalization dependent on the morphology

In order to lowercase the first word of a sentence, the Wolof morphological analyzer described in Dione (2012) has been extended with the function in (35). Decapitalization applies then for those words marked by the tokenizer with caret, indicating that they were found at the beginning of the sentence. This function is designed such that it only handles words marked as such, ignoring other words found somewhere else. It also removes the caret symbol after decapitalization is performed by the *downcase* term.

```
(35)  define MarkDowncased(X) [ [\"^"]* .o. X ]
                              | [ X .o. [ 0:"^" ( downcase ) ?* ]];
```

The *downcase* term denotes the inverse of the term *upcase*, as defined in (36).

```
(36)  define upcase \> [ \> A:a|B:b|C:c|D:d|E:e|F:f|G:g|... ];
```

The term in (36) contains a number of pairs and represents the mapping of all uppercase strings to the corresponding lowercase strings. It consists of ordered pairs *<A,a>* of symbols *A:a*, where *A* is the upper-side symbol and *a* is the lower-side symbol. The upper language is the infinite language of uppercase strings, the lower language contains all the lowercase strings, and the term itself is a mapping that preserves the word. Thus, if `upcase` contains *<A, a>*, the inverse relation `upcase.i` contains *<a, A>*.

In a final stage (37), the function `MarkDowncased` is applied to the main Wolof transducer `WolMorph` which represents all those words handled by the Wolof morphology.

```
(37)  read regex MarkDowncased (WolMorph);
```

Using the normalizing transducer, the first word of the sentence in (14) will be lowercased, as illustrated by the tree on the left side of Figure 5. In contrast, the tree on the right side, shows how the spelling of proper names like *Móodu* in the sentence *Móodu dem* 'Móodu left' is preserved.

5 Evaluation

The tokenizer can be evaluated in the context of the standard Wolof LFG parser (Dione 2014a) which makes use of the tokenizer to annotate free text. The performance of the parser was measured on unseen natural text data consisting of 2354 sentences randomly selected from Cissé (1994), Garros (1997) and Ba (2007). The parser was evaluated in terms of coverage and parsing quality. Coverage indicates whether the parser yields

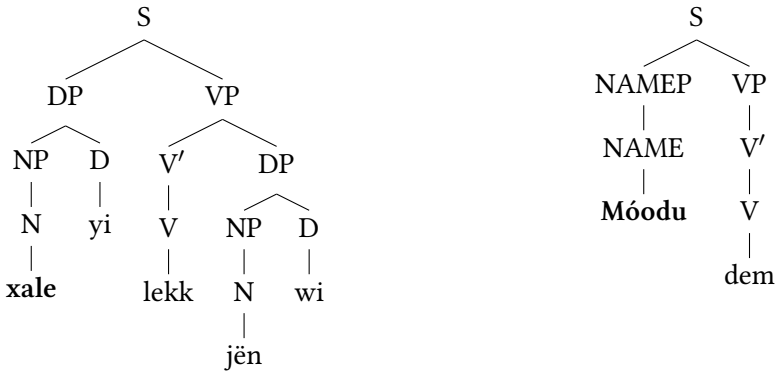


Figure 5: Normalizing the first word of sentence (14)

complete parses or not. Accordingly, the evaluation results given in Dione (2014a) reported that the Wolof parser could find a complete parse for 1712 of the test sentences (i.e. 72.72% coverage for complete parses).

For a direct evaluation of the tokenizer, 350 out of the 642 sentences that could not be parsed were randomly selected to determine whether parsing failure was due to erroneous tokenization. For 330 of them, this was not the case. Among the twenty that failed due to tokenization errors, ten are due to inappropriate treatment of clitics, mostly caused by vowel assimilation or complex derivation (e.g. reduplication); five contain multi-word units, including names; two sentences contain all uppercase strings like *BUKKEEK «PERIGAM» BU XONQ* 'Hyena and its red «wig»' which also involve issues related to clitics and quotes; the rest consists of tokenization errors due to the use of symbols and foreign language material as well as a mixture of the issues discussed so far.

The problem of the multiword units is difficult to address without good lists of person, place, organization and product names. In many cases, the tokenization problems are caused by different issues. For instance, the clitic transducer erroneously inserted a space between *ja* and *ag* in the string (38), considering both as determiners (i.e. *ja* 'the' and *ag* 'a'), which might be correct in some contexts. In (38), however, *Ɉaag* is a last name which can be treated either as an individual token or as a part of a multiword expression, but not segmented into two strings. Adding this named entity to the list of identified MWEs would help to avoid this kind of problems.

(38) Sàmba Jaag

Likewise, issues related to clitics are complex and need to be addressed in future work on the tokenizer.

6 Conclusion

This paper has presented a finite-state transducer for tokenizing normal Wolof text. It has shown that the design of such a preprocessing tool involves non-trivial issues related to the treatment of clitics, multiword expressions and text normalization. Accordingly, sophisticated techniques covering these issues have been integrated into the tokenization model. Also, the paper has explained how the different preprocessing components interact with each other and how tokenization and normalization are closely connected to and sometimes dependent on morphological analysis.

However, as this paper acknowledges, there are open tokenization issues that need to be addressed in future work. This includes, for instance, better handling of clitics by integrating sophisticated techniques to control ambiguity caused by the guessing mechanisms. Similarly, robust corpus-based approaches to multiword extraction need to be combined with tokenization. Finally, future work on text normalization include issues related to capitalization and haplology (Forst and Kaplan 2006).

7 Acknowledgements

This research has received support from the EC under FP7, Marie Curie Actions SP3-People-ITN 238405 (CLARA). I thank Paul Meurer for helping me integrate the normalizer into the morphological analyzer.

References

Attia, Mohammed A. (2007). "Arabic Tokenization System". In: *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Association for Computational Linguistics, pp. 65–72.

– (2008). "Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation". PhD thesis. University of Manchester.

Ba, Mariyaama (2007). *Bataaxal bu gudde nii*. Nouvelles Editions Africaines du Sénégal (NEAS), p. 182.

Beesley, Kenneth R. and Lauri Karttunen (2003). *Finite State Morphology*. Stanford, CA: Center for the Study of Language and Information.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer (2002). "The Parallel Grammar Project". In: *Proceedings of the COLING2002, Workshop on Grammar Engineering and Evaluation*. Vol. 15. Association for Computational Linguistics, pp. 1–7.

Church, Eric D. (1981). *Le système verbal du wolof*. Faculté des Lettres et Sciences Humaines (FLSH), Université de Dakar.

Cissé, Mamadou (1994). *Contes wolof modernes*. Paris: L'Harmattan.

Dione, Cheikh M. Bamba (2012). "A Morphological Analyzer For Wolof Using Finite-State Techniques". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: ELRA.

– (2014a). "Formal and Computational Aspects of Wolof Morphosyntax in Lexical Functional Grammar". PhD thesis. University of Bergen, Norway.

– (2014b). "LFG parse disambiguation for Wolof". In: *Journal of Language Modelling* 2.1, pp. 105–165.

– (2014c). "Pruning the Search Space of the Wolof LFG Grammar Using a Probabilistic and a Constraint Grammar Parser". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 2863–2870.

Forst, Martin and Ronald M. Kaplan (2006). "The importance of precise tokenizing for deep grammars". In: *Proceedings of the Language Resources and Evaluation Conference (LREC'06)*. Genoa, Italy.

Garros, Nataali Dominik, ed. (1997). *Bukkeek "perigam" bu xonq: teeñ yi*. Dr. Moren ak mbootayu "xale dimbale xale". Translated from French to Wolof by Momar Touré. Dakar: SIL / Paris: EDICEF.

Ka, Omar (1994). *Wolof Phonology and Morphology*. Lanham, Maryland: University Press of America.

Kaplan, Ronald M. (2005). "A method for tokenizing text". In: *Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday*. Stanford, CA: CSLI Publications.

Kaplan, Ronald M., John T. Maxwell III, Tracy Holloway King, and Richard Crouch (2004). "Integrating Finite-State Technology with Deep LFG Grammars". In: *Proceedings of the ESSLLI'04 Workshop on Combining Shallow and Deep Processing for NLP*.

Karlsson, Fred (1990). "Constraint Grammar as a Framework for Parsing Running Text". In: *Proceedings of the 13th Conference on Computational Linguistics*. Vol. 3. ACL. Helsinki, pp. 168–173.

McLaughlin, Fiona (2010). "Noun classification in Wolof: When affixes are not renewed". In: *Studies in African Linguistics* 26.1, pp. 1–28.

Oflazer, Kemal, Özlem Çetinoğlu, and Bilge Say (2004). "Integrating morphology with multi-word expression processing in Turkish". In: *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*. Association for Computational Linguistics, pp. 64–71.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger (2002). "Multiword Expressions: A Pain in the Neck for NLP". In: *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*. Vol. 2276. Lecture Notes in Computer Science. Springer, pp. 1–15.

Wu, Dekai and Pascale Fung (1994). "Improving Chinese tokenization with linguistic filters on statistical lexical acquisition". In: *Proceedings of the Fourth Conference on Applied Natural Language Processing*. Association for Computational Linguistics, pp. 180–181.

Yang, Christopher C. and Kar Wing Li (2005). "A heuristic method based on a statistical approach for Chinese text segmentation". In: *Journal of the American Society for Information Science and Technology* 56.13, pp. 1438–1447.

Syntactic discontinuities in Latin — A treebank-based study

Dag Haug

Abstract. Syntactic discontinuities are very frequent in classical Latin and yet this data was never considered in debates on how expressive grammar formalisms need to be to capture natural languages. In this paper I show with treebank data that Latin frequently displays syntactic discontinuities that cannot be captured in standard mildly context-sensitive frameworks such as Tree-Adjoining Grammars or Combinatory Categorial Grammars. I then argue that there is no principled bound on Latin discontinuities but that they display a broadly Zipfian distribution where frequency drops quickly for the more complex patterns. Lexical-Functional Grammar can capture these discontinuities in a way that closely reflects their complexity and frequency distributions.

1 Introduction

Classical Latin, like classical Greek, is famous for its tolerance of syntactic discontinuities. One example is shown in (1).

(1) *quis*$_1$ *multa*$_2$ *gracilis*$_1$ *te* *puer*$_1$ *in* *rosa*$_2$
who.NOM much.ABL slender.NOM you.ACC boy.NOM in rose.ABL
perfusus *liquidis*$_3$ *urget* *odoribus*$_3$ *grato*$_4$, *Pyrrha,*
drenched.NOM liquid.ABL press.3SG.PRES scents.ABL delightful.ABL Pyrrha
sub *antro*$_4$
in cave.ABL
'What slender boy, drenched with perfumes, presses you on a bed of roses, Pyrrha, under the delightful cave?' (Horace, Carmina 1.5)

This example features no less than four discontinuous noun phrases, as indicated with subscript indices on the words. The syntactic dependencies inside these NPs are marked with agreement in case (and number and gender, not shown in the glossing), but not with word order.

Discontinuous NPs are in fact attested in Latin up to the twentieth century, as in (2) from Dyvik (1968).

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1337.

(2) *Mihi, Paulo, nullus est terror*
me.DAT Paul.DAT none.NOM is fear.NOM
'I, Paul, have no fear.'

Examples such as (2) are reminiscent of quantifier float, a type of discontinuity which is found even in highly configurational languages such as English. While (2) is in fact the only discontinuous NP in Dyvik (1968), the focus of this article is on the classical stage of Latin, where many other types of discontinuities are attested, as shown already in (1).

At the same time as Dyvik (1968) was composed, linguists discussed whether natural languages are context-free. The debate was sparked by the definition of the Chomsky hierarchy in Chomsky (1956), which raised the question whether natural languages could be described by context-free grammars. This was an open question throughout the 1960s and 70s and it was not until the 1980s that the question was settled (in the negative).[1]

The classical languages played little role in this discussion. Occasionally, some Latin examples were cited — for example, Ross (1967/1986, p. 74) used (1) to illustrate *scrambling*. There is no obvious characterization of the elements that can intervene between the different parts of the NPs in this example, and if we assume that there is no theoretical upper bound on the intervening material, it looks like it could be possible to construct an argument for the non-contextfreeness of natural language based on such examples. Of course, assuming that there is no upper bound is a leap of faith that could never be truly justified — but in that respect, Latin is not really different from English. The common claim that finite state automata cannot model center-embedding also depends on there being no theoretical upper bound on the level of embedding, and neither in English nor in Latin can we observe infinite embeddings. In fact, corpus studies suggest that the practical upper bound on levels of center embeddings is as low as three. So the main reason for using a context-free grammar to deal with center-embedding is theoretical simplicity and elegance, as pointed out by Harris (1957): "If we were to insist on a finite language, we would have to include in our grammar several highly arbitrary and numerical conditions — saying, for example, that in a given position there are not more than three occurrences of *and* between N".

Similar considerations would apply to Latin discontinuities. However, to the extent that Latin examples featured in the scholarly discussion, scholars did not object to them because their unboundedness could not be demonstrated. Instead, Pullum (1982) argued that (1) comes from the poet Horace, who "is noted for stretching tendencies in the living Latin language beyond all grammatical limits". And so no one attempted to build an argument based on classical data. Another reason for suspicion, no doubt, was the lack of hard facts concerning the extent of syntactic discontinuity in a dead

1 See Pullum (1986) for a fascinating account of the debate.

language like Latin. The Latin grammatical tradition has been content to establish that word order is 'generally free' (not just in poetry, but also in prose) and to investigate the stylistic usage of discontinuity. Moreover, word order data from Greek and Latin used not to be very accessible. As shown in Haug (2015), it used to be the case that scholars could not even agree on the frequencies of basic word orders in Ancient Greek – never mind providing an account of them.

2 So how complex is natural language really?

Research in the Generalized Phrase Structure Grammar (GPSG) framework in the early 1980s was to a large extent motivated by the desire to keep the complexity of the formalism low and develop context-free analyses of seemingly non-context free phenomena such as long distance dependencies (Gazdar 1981). That program imploded when Shieber (1985) and Culy (1985) showed that there are phenomena in natural language that cannot be captured in a context-free grammar. There were two main responses to this discovery: either one tried to extend context-free formalisms as little as possible while achieving coverage of demonstrably non-context free phenomena such as the cross-serial dependencies from Dutch and Swiss German discussed in Shieber (1985), leading to so-called mildly context-sensitive formalisms such as (Lexicalized) Tree Adjoining Grammar (LTAG) and Combinatory Categorial Grammar (CCG); or one gave up (almost) completely on the concern about weak generative capacity, as in Lexical Functional Grammar (LFG) and Head Driven Phrase Structure Grammar (HPSG). A natural question to ask, then, is "Who was right?". Is it possible to keep the algorithmic complexity of the parsing problem low while maintaining good coverage of the data, as measured in modern treebanks?

Answering that question requires a detour into dependency grammar, since most treebanks these days – and in particular the Latin ones that we will look at here – are based on dependencies rather than phrase structures or CCG derivations. Fortunately, there are formal results that relate the complexity of formalisms like CFGs, LTAGs and CCGs to that of dependency grammars with various restrictions on non-projective (i.e. discontinuous) dependencies.

2.1 Measuring discontinuity in a dependency treebank

In order to study discontinuities in dependency trees, we need to introduce some terminology. The *projection* of a node in a dependency tree is its yield, i.e. the set of nodes in the transitive, reflexive closure of dominance, arranged in linear order. A *gap* is a discontinuity in a projection, and the *gap degree* of a node is the number of gaps in its projection. An equivalent measure is the block degree, i.e. the number of continuous *blocks* in the projection of a node, which will always be the gap degree + 1. Consider the dependency tree in Figure 1. The gap degree of *mihi* is 0, for its projection [*mihi, Paulo*] is uninterrupted. By contrast, the gap degree of *terror* is 1, for its projection [*nullus, terror*] is interrupted by *est*. Alternatively, we may say that *terror* has block

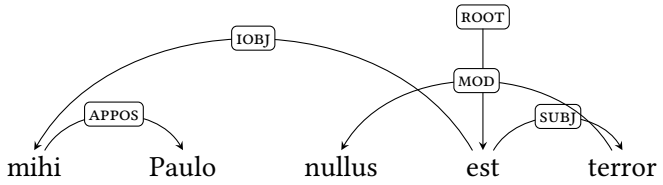


Figure 1: Dependency tree for (2)

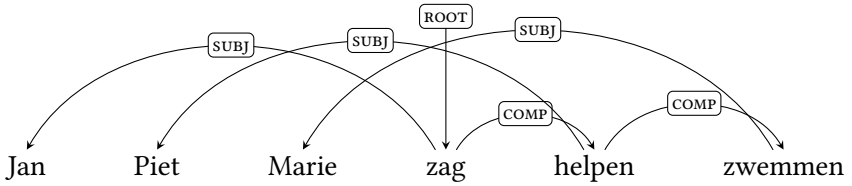


Figure 2: Dependency tree for (4)

degree 2, for it consists of the two blocks [*nullus*] and [*terror*]. Finally, we note that we may also talk about the gap degree of a dependency tree, which is defined as the highest gap degree among its nodes.

For our purposes, it will also be useful to study gap *depth*, which we define as in (3).

(3) A node d in the projection of r *introduces a discontinuity* in r iff d is in a different block b from r and there is no node in b that dominates d. The depth of the gap introduced by d is the number of edges between d and r. The *gap depth* of r is the maximum depth of a node that introduces a discontinuity in r.

In Figure 1, the gap depth of *terror* is 1, as the discontinuity is introduced by its direct dependent *nullus*. Let us now look at a deeper gap in a classic example of cross-serial dependencies in Dutch.

(4) *(…dat) Jan Piet Marie zag helpen zwemmen*
that Jan Piet Marie see-PST help.INF swim.INF
'(…that) Jan saw Piet help Marie swim.'

The nodes *helpen* and *zwemmen* both have gap degree 1. The projection of *helpen* has the two blocks {*Piet, Marie*} and {*helpen, zwemmen*}. Both *Piet* and *Marie* introduce discontinuities in the projection of *helpen*, since neither dominates the other. The depth

of those discontinuities are 1 and 2 respectively and hence the gap depth of *helpen* is 2. Thus the gap depth captures the fact that not only is *helpen* discontinuous, but it also dominates a discontinuous dependent without resolving the discontinuity. Intuitively, then, gap depth captures embedding of discontinuities e.g. in long-distance extraction, which is generally thought to be associated with human processing difficulty (see e.g. Gibson 2000). Gap depth has to my knowledge never been considered in measures of non-projectivity in dependency treebanks such as Kuhlmann and Nivre (2006), Havelka (2007) or Maier and Lichte (2011), but we will see that corpus evidence suggests this measure is useful.

2.2 Dependency structures and other grammatical formalisms

While most modern treebanks are based on dependencies, most grammatical theories are not. One early and fairly well-known result connecting dependency grammar to other grammatical formalisms is due to Gaifman (1965) and shows that projective dependency grammars, i.e. dependency grammars that allow no discontinuities, are weakly equivalent to context-free grammars. However, since the focus here is precisely on discontinuities, that result is of little value for us.

Multiple context-free grammars (MCFGs), also known as linear context-free rewriting systems, have emerged as a powerful tool to study complexity questions in the range of the Chomsky hierarchy between context-free grammars and full-blown context-sensitive grammars. Kuhlmann (2013) has established connections between dependency grammars and MCFGs which yield a close correspondence between the non-projectivity of the dependency trees admitted by a grammar on the one hand, and the parsing complexity of the grammar on the other. In the following, we briefly review these results as a background for what follows.

The MCFG formalism is a generalization of CFG which retains ordinary CFG productions for the expression of categorial structure, but uses explicit *yield functions* to compute the yield of the mother node from the yields of the daughters. In an ordinary CFG, yield computation is conflated with category formation: a rule such as DP $\rightarrow$ D NP says both that the category DP is formed of a D and an NP, and that the yield of the resulting DP is formed by concatenating the yields of D and NP. In effect, then, a CFG can be seen as an MCFG with concatenation as the only yield function.[2]

To allow for greater expressivity, MCFG allows yields to be *tuples* of strings. For example, we may want to say that the yield of DP is a pair (2-tuple) consisting of the yields of D and NP. This pair will then be the input to further yield functions that apply to productions with DP on the right-hand side. More generally, we may allow yields to be n-tuples of strings.

For our purposes, it is important to note that there is a close correspondence between yield components in an MCFG and blocks in a corresponding dependency structure. We can extract MCFG rules from dependency trees, as shown in Kuhlmann (2013),

2 See Clark (2014) for an accessible introduction for linguists.

APPOS $\rightarrow f()$	$f = \langle \textit{Paulo} \rangle$
MOD $\rightarrow g()$	$g = \langle \textit{nullus} \rangle$
IOBJ $\rightarrow h($APPOS$)$	$h = \langle \textit{mihi}\ x_1 \rangle$
SUBJ $\rightarrow i($MOD$)$	$i = \langle x_1, \textit{terror} \rangle$
ROOT $\rightarrow j($IOBJ SUBJ$)$	$j = \langle x_1 y_1\ \textit{est}\ y_2 \rangle$

Table 1: Rules extracted from the tree in Figure 1

where a formal exposition is given. Here I just provide an intuitive understanding of how the tree in Figure 1 gives rise to the rules in Table 1.

Looking at *Paulo* in Figure 1 we see that it has no dependents, hence the right-hand side of the first rule is a constant function which fixes the yield to the string *Paulo*, and similarly for *nullus*. For *mihi*, things are a bit more interesting: it takes an APPOS argument, and hence its yield depends on the yield of that argument. Concretely, the yield of the node *mihi* is computed by concatenating the string *mihi* with the yield of the APPOS argument, which is represented with x_1 according to the convention that we use x for the yield of the first argument and y for the yield of the second argument, and subscript those variables with an index referring to components of the yield. In this case, the yield of APPOS has only one component, so we use x_1. Also *terror* takes an argument, a MOD, but in this case, the resulting yield has two components, one consisting of the yield of the MOD and one consisting of the string *terror*. Finally, the verb takes two arguments, SUBJ and IOBJ. The yield is constructed by concatenating the yield of the IOBJ(i.e. x_1), the first component of the SUBJ(i.e. y_1), the string *est*, and the second component of SUBJ(y_2).

For our purposes, the primary interest of this construction lies in the fact that it provides a link between dependency treebanks and the required expressivity of corresponding grammars, as investigated in Kuhlmann (2013). On the one hand, the yield components correspond directly to blocks found in the treebanks. And on the other hand, the complexity of an MCFG grammar is easily read off the yield functions: The parsing complexity of a yield function equals the sum of the number of components in its input and output yields. For example, the parsing complexity of j in Table 1 is 4, as its two inputs have 1 and 2 component yields and it produces a 1 component yield. This yields a two-dimensional complexity hierarchy, as the complexity depends both on the number of arguments and the number of yield components of these arguments. In the presence of only *wellnested* discontinuities, we actually get a simple complexity hierarchy because any wellnested MCFG can be binarized without increasing the gap degree. A wellnested discontinuity is one whose projection does not interleave with

another non-overlapping projection.[3] (5) gives an example of an illnested discontinuity from Latin poetry.

(5) *aurea* *purpuream* *subnectit* *fibula* *vestem*
golden.NOM purple.ACC bound clasp.NOM cloak.ACC
'a golden clasp bound her purple cloak' (Vergil, Aeneid 4.139)

The projections of the subject and the object do not overlap (neither dominates the other), but they interleave, producing an illnested discontinuity. MCFGs that generate only wellnested dependencies are called wellnested MCFGs. Since they can be binarized without increasing the gap degree, their parsing complexity is uniquely determined by their gap degree. We refer to an MCFG where no argument has more than k components in its yield as a k-MCFG.

There are several results linking linguistically motivated grammatical formalisms to MCFGs. For example, TAG is weakly equivalent to a wellnested 2-MCFG. The same result applies to 'classical' CCG and linear indexed grammars (Aho 1968), since those formalisms are weakly equivalent to TAG. However, modern lexicalized CCG (i.e. the current version where (restrictions on) the combinators are not grammar-specific but all linguistic variation is captured in the lexicon) is known to be strictly less powerful than TAG (Kuhlmann, Koller, et al. 2015).

The equivalence between wellnested 2-MCFGs and established grammatical formalisms takes on significance in the light of empirical investigations on dependency treebanks. For example, Kuhlmann (2013) shows that by restricting ourselves to wellnested trees of gap degree at most 1, i.e. trees describable by a wellnested 2-MCFG, we lose only between 0.1% (Arabic) and 0.9% (Turkish) of the trees in the CoNLL 2006 treebanks. This suggests that formalisms with the power of TAG are adequate for natural languages. Similar results have been reported by others and will also be shown below for the Universal Dependencies treebanks. But we will also see that Latin behaves in a crucially different way.

2.3 Complexity in LFG

Any LFG grammar that determines an upper bound n on the number of c-structure nodes corresponding to a given f-structure (a so-called 'finite copy LFG') can be translated into a weakly equivalent MCFG. This gives us polynomial time parsing, because parsing with a (wellnested) k-MCFG can be done in time $\mathcal{O}(n^{3k})$. But in the general case, parsing with an LFG grammar is NP-complete, as can be shown with a straightforward reduction from the 3SAT problem, i.e. the problem of determining whether a formula of propositional logic in conjunctive normal form where each clause is limited to at most three literals is satisfiable: we use c-structure rules to make sure each clause

3 For a formal definition, see e.g. Kuhlmann (2013, p. 377).

contains at least one true literal and use the f-structure to keep track of the assignment of truth values across clauses.[4]

It is worth pointing out that the universal recognition problem for MCFGs is also NP-complete because, although any given MCFG is a k-MCFG, MCFG as a formalism does not bound that k. Put in other words, the difference between MCFGs and LFGs is that for any given (finite) instance of the 3SAT problem with n clauses, we can construct an MCFG that solves it, whereas we can write a general LFG that can solve any instance of the 3SAT problem.

If we think of the relations between different instances of the same literals in a 3SAT problem as analogues to discontinuous dependencies in linguistics, this means an LFG grammar can deal with an unbounded number of discontinuous dependencies across unbounded distances. We can ask ourselves whether there is any need for the expressivity that LFG gives us. As we will see in section 3.1, the answer is from one point of view negative: we can get extremely good coverage on existing dependency treebanks with a relatively low bound on the discontinuities. Nevertheless, it is worth making the point that the extra expressivity provides for extra linguistic insight. We will now show that this point holds even as we move up the complexity ladder from NP-complete to undecidable.

Undecidability was not a property of LFG originally. While unification grammars in general are Turing-equivalent and hence have an undecidable parsing problem, Kaplan and Bresnan (1982) avoided undecidability by restricting valid derivations as in (6).

(6) A c-structure derivation is valid if and only if no category appears twice in a nonbranching dominance chain, no nonterminal exhaustively dominates an optionality ϵ, and at least one lexical item or controlled e appears between two optionality ϵ's derived by the same rule element.

By disallowing nonbranching dominance chains, this constraint ensures that for any string the size and number of c-structure derivations is bounded as a function of the length of the string. The constraint seems well-motivated: after all, what could be the linguistic motivation for derivations in which e.g. some NP dominates another NP in a nonbranching structure?

As it turns out, such structures *can* be motivated. In Bresnan, Kaplan, et al. (1982), it was argued that cross-serial dependencies in Dutch cannot be given a linguistically motivated analysis in a context-free grammar. Instead, the authors proposed to give the sentence in (4) the c-structure in Figure 3.

This c-structure does not directly capture the object relation between *Piet* and *zag* or between *Marie* and *helpen*. Instead, the relationship is captured with functional annotations on the VP and $\bar{\text{V}}$ nodes which 'match up' the two branches in the f-structure and give the correct grammatical relations. So, *Marie* is the object of *helpen* by virtue of

4 See for example Francez and Wintner (2012, pp. 241–243) for details of the construction.

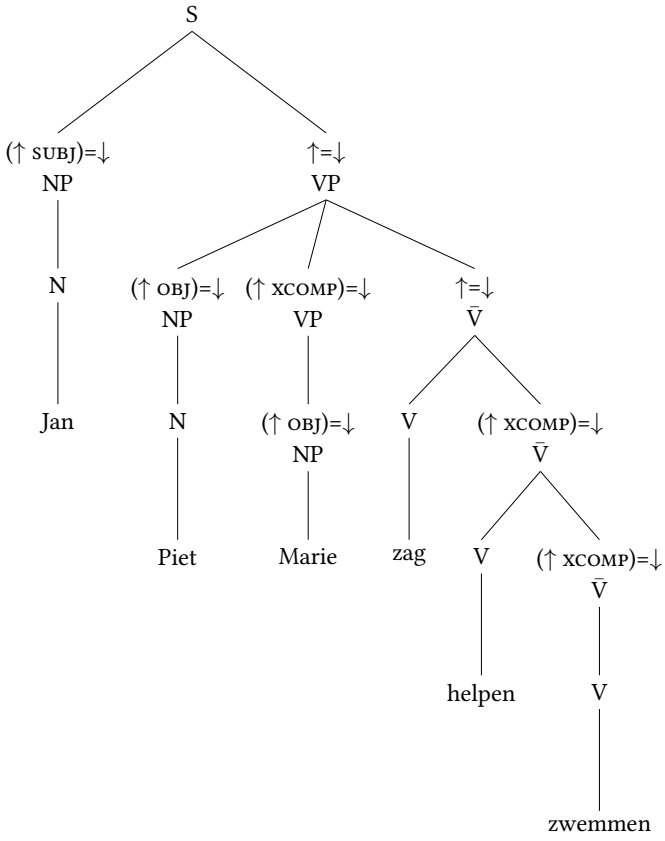


Figure 3: C-structure of (4)

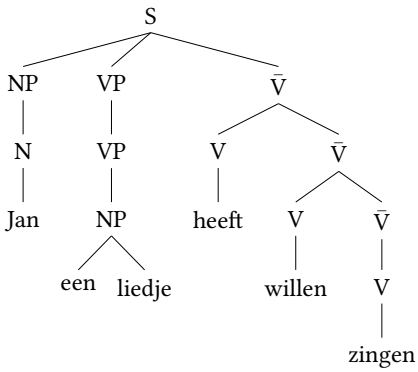


Figure 4: C-structure of (7)

being embedded under the same number of VP nodes as *helpen* is under V̄ nodes. This kind of analysis is based on what Maxwell and Kaplan (1996) call 'zipper unification'.

However, as pointed out by Johnson (1986), this analysis actually leads to non-branching dominance chains in cases where intermediate verbs in the structure are intransitive, as in (7) with the c-structure in Figure 4.

(7) *(...dat) Jan een liedje heeft willen zingen*
that Jan a song has want.INF sing.INF
'(...that) Jan has wanted to sing a song.'

So, if we want to keep the analysis from Bresnan, Kaplan, et al. (1982) we must give up the offline parsability constraint and hence the decidability of the LFG formalism. On the other hand, an alternative analysis was also proposed (Zaenen and Kaplan 1995), where NPs inside VP get the functional uncertainty annotation $(\uparrow \text{XCOMP}^* \text{OBJ}) = \downarrow$, rather than just $(\uparrow \text{OBJ}) = \downarrow$. From a linguistic point of view, there are several problems with this analysis: First, it is unclear how we can ever provide a principled structure-function mapping if we allow non-local GF assignments like this. And second, in order to capture the word order facts, we need complex f-precedence constraints.

And in fact, what happened in this case is that the analysis of Bresnan, Kaplan, et al. (1982) is still well-known and cited, whereas the alternative analysis based on non-local GF assignment and functional precedence is more or less forgotten. Both the first and the second edition of Bresnan's LFG textbook (Bresnan 2001; Bresnan, Asudeh, et al. 2015) include exercises that ask the student to reproduce Bresnan, Kaplan, et al. (1982) — even if a generalization of this analysis to intransitive verbs (not used in the exercise) would not even be LFG as defined in Kaplan and Bresnan (1982). In other

words, while 'intuitive' is a subjective notion, history lends some justification to the claim that the original analysis is more intuitive than the later one.

There are some lessons we can draw from this. First, the ban on nonbranching dominance chains looks stipulative: it can be removed from the definition of LFG without changing anything else, albeit at the cost of undecidability. An analysis like that in Figures 3 and 4 does not 'feel' substantially un-LFG-like. Second, the original analysis seems linguistically more informative than the alternative in that it captures the word order generalizations in an intuitive way while preserving locality of GF assignment. Again, this is subjective, but the fact that the analysis gets cited and is used in textbooks shows that the intuition is widespread.

Taken together, these two observations suggest that a more expressive grammatical formalism can lead to more linguistically adequate analyses — even if those analyses do not actually exploit that expressivity in a crucial way. In our case, the problem with unary branching dominance chains is that there will be no upper bound on the length of the unary VP chain in Figure 4. But chains of unbounded length are of course not crucial to the analysis. We only need VP–VP chains of a length corresponding to the number of consecutive intransitive verbs in the $\bar{\text{V}}$-chain. For practical purposes, 5 will be more than sufficient. And even from a theoretical perspective, it is not clear that banning any category α from dominating five instances of α in a nonbranching dominance chain is any more objectionable than banning it from dominating a single instance, as Bresnan and Kaplan did with (6).

3 Empirical investigation

3.1 Quantitative data

Let us now have a look at how discontinuities actually distribute in Latin treebanks. To be able to compare across languages we use the Universal Dependencies (UD) corpora,[5] in particular the version 2 release. This dataset contains three Latin treebanks, the Perseus treebank (Bamman and Crane 2011), the PROIEL treebank (Haug and Jøhndal 2008) and the Index Thomisticus Treebank (Martens and Passarotti 2014).

Table 2 shows the distribution gap degree and depth across all languages in the UD corpora.[6] As we can see, the vast majority of edges, 97.3%, are projective. Still, this means that the number of non-projective edges is high enough that we need to be able to deal with them in parsing. But at least from a practical standpoint, we can ignore everything but the simplest type of gap: restricting ourselves to edges of gap degree and depth ≤ 1 yields a coverage of 99.7%.

When we get to Latin, the picture is different. First of all, the number of simple (degree 1) non-projectivities is much higher: 9.1% of edges. More interesting is the fact

5 http://universaldependencies.org/

6 There are a few outliers of degree or depth > 3 that are not shown in the tables.

Gap degree	Gap depth			
	0	1	2	3
0	2572961 (97.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
1	0 (0.0%)	63729 (2.4%)	4856 (0.2%)	616 (0.0%)
2	0 (0.0%)	1584 (0.1%)	1095 (0.0%)	165 (0.0%)
3	0 (0.0%)	44 (0.0%)	56 (0.0%)	25 (0.0%)

Table 2: Gap degree and depth in the UD 2.0 corpora

Gap degree	Gap depth			
	0	1	2	3
0	60430 (90.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
1	0 (0.0%)	5556 (8.3%)	496 (0.7%)	61 (0.1%)
2	0 (0.0%)	134 (0.2%)	123 (0.2%)	17 (0.0%)
3	0 (0.0%)	5 (0.0%)	5 (0.0%)	1 (0.0%)
4	0 (0.0%)	0 (0.0%)	1 (0.0%)	0 (0.0%)
9	0 (0.0%)	0 (0.0%)	1 (0.0%)	1 (0.0%)

Table 3: Gap degree and depth in the UD 2.0 Latin-PROIEL treebank

that 0.4% of edges have gap degree 2 and thus reflect dependencies that cannot be captured in a TAG (or a forteriori, in a CCG).

This becomes clearer if we think about tree coverage, as shown in Table 4 for a select number of treebanks.[7] Here we see that by restricting ourselves to trees where the highest gap degree is 1, we lose 1.8% of the trees in the PROIEL treebank, compared to zero loss in the Norwegian Bokmål treebank and 0.3% loss in the Czech treebank. Overall in the UD treebanks, 0.6% of trees contain an edge of gap degree 2, but it is worth pointing out that almost three quarters of these trees are found in one of the Ancient Greek and Latin treebanks, which only make up roughly a tenth of the trees. So there clearly is something special about these languages.

Finally, we look at the illnestedness numbers in Table 5. As has been observed several times in the literature, illnestedness is a strong constraint on discontinuities in most languages. We see that this constraint is strong also in the PROIEL corpus of Latin (and Greek), but not in the Perseus corpora. As with non-projective dependencies in general, this is likely due to due to the large portions of poetry in this treebank. In (5) we saw an example of an illnested dependency from Vergil. And this was in fact no accident,

7 What I have listed as Anc.Gr.-Perseus and Latin-Perseus appear in UD as simply Ancient Greek and Latin, since they were the first treebanks for these languages. These treebanks generally have a higher degree of non-projectivity because they consist mostly of poetry.

	0	1	2	3
Anc.Gr.-Perseus	4738 (37.6%)	6983 (55.4%)	833 (6.6%)	46 (0.4%)
Anc.Gr.-PROIEL	9823 (61.9%)	5515 (34.8%)	493 (3.1%)	34 (0.2%)
Czech	3480 (87.9%)	468 (11.8%)	11 (0.3%)	0 (0.0%)
Latin-Perseus	793 (59.5%)	511 (38.3%)	27 (2.0%)	2 (0.2%)
Latin-ITTB	10367 (62.9%)	5805 (35.2%)	310 (1.9%)	9 (0.1%)
Latin-PROIEL	11213 (73.2%)	3844 (25.1%)	254 (1.7%)	11 (0.1%)
Norw.-Bokmaal	1173 (92.5%)	95 (7.5%)	0 (0.0%)	0 (0.0%)
All treebanks	353194 (87.2%)	49151 (12.1%)	2572 (0.6%)	121 (0.00%)

Table 4: Trees by gap degree in selected UD treebanks

	Illnested	Wellnested
Ancient_Greek-Perseus	1.5%	98.5%
Ancient_Greek-PROIEL	0.3%	99.7%
Czech	0.1%	99.9%
Latin-Perseus	3.8%	96.2%
Latin-ITTB	0.2%	99.8%
Latin-PROIEL	0.2%	99.8%
Norwegian-Bokmaal	0.1%	99.9%
Sum	0.1%	99.9%

Table 5: Wellnestedness

but a so-called 'golden line', a rhetorical pattern first discovered by Edward Burles in 1652: "If the Verse does consist of two Adjectives, two Substantives and a Verb only, the first Adjective agreeing with the first Substantive, the second with the second, and the Verb placed in the midst, it is called a Golden Verse." It is not clear whether Latin poets in fact preferred illnested dependencies for their own sake, or whether their frequency results from other, conspiring factors. Whatever the motivation, it is interesting that the poets regularly produced these illnested structures which are so rare in prose.

The numbers reported in this section are based on data converted to Universal Dependencies. To my knowledge there is no in-depth study of discontinuity based on the original Perseus or PROIEL data for Latin, but there is a study on Greek (Mambrini and Passarotti 2013), which finds only 25.2% projective trees, compared to 37.6% in the UD version of the same treebank. It should be noted that the UD conversion only includes a subset of the original treebank due to conversion problems. One possibility is that the conversion script was particularly likely to fail on non-projective structures,

which would explain why the projectivity rate is higher in the converted UD data. The illnestedness degree is also lower, at 1.5% versus 2.6% in the original version.

3.2 Examples

Let us now have a closer look at some examples of the discontinuities we find in Latin. An important first observation is that a large number of them arise from second-position clitics, which normally appear after the first prosodic word, even if that breaks up a syntactic constituent. The frequency of this phenomenon contributes to the number of gap degree 2 trees, since it is then enough to have one other gap resulting from some other process. An example of this is (8).

(8) *eo* *autem* *die* *credo* *aliquid* *actum* *in* *senatu*
this.ABL but day.ABL I.believe something.ACC done.ACC in senate.ABL
'But I believe that something will be done in the senate today.' (Cic. Att. 5.5.1)

In this case, we have a normal long distance dependency, where *eo die* has been displaced out the embedded clause *aliquid actum in senatu*, resulting in one gap. When the clitic then lands inside the fronted constituent, we get a second gap. Such examples are controversial as illustrations of the syntactic complexity of a language, since it is not clear to what extent clitic positioning in Latin is syntactically conditioned: prosodic factors are clearly also important. From a parsing perspective, however, we need to have some way of dealing with clitics, so a more reasonable objection may be that the set of clitic strings is finite, i.e. there is only a finite number clitics and licit combinations of clitics that can occur in the position of *autem* in (8). Therefore, we can deal with them without using the full power of a formalism that can derive syntactic discontinuities.[8]

However, trees of gap degree 2 are by no means restricted to those where clitics account for one of the gaps. Example (9) shows a discontinuous NP *multa …genera ferarum*, with an extraposed relative clause.

(9) *Multa* *que* *in* *ea* *genera* *ferarum* *nasci* *constat*
many.ACC and in it.ABL kinds.ACC beasts.GEN be born it.is.certain
quae *reliquis* *in* *locis* *visa* *non* *sint*
which.NOM other.ABL in places.ABL seen.NOM not are.SBJV
'It is certain that many kinds of beasts are born in it which have not been seen in other places' (Caes. Gal. 6.25.5)

(10) shows another example, where we get gap degree 2 because the genitive is displaced from its head noun at the same time as the wh-word *quantam* is fronted alone.

8 An approach based on MCFGs can still be more perspicuous and insightful from a linguistic point of view, see Goldstein and Haug (2016). Even so, it is likely that such a grammar could be 'compiled' to a computationally more tractable grammar by exploiting the finiteness of the set of clitic strings.

(10) *quantam* *porro* *mihi* *exspectationem* *dedisti* *convivi*
how large.ACC besides me.DAT expectation.ACC give.2S.PRF this.GEN
istius *ἀσελγοῦς*
guest.GEN wanton.GEN
'Besides, how large expectations you gave me about this wanton guest!' (Cic. Att. 2.12.2)

Example (11) shows another common pattern, where one of the gaps results from a 'postponed' coordination.

(11) *Munitis* *castris* *duas* *ibi* *legiones* *reliquit* *et*
fortified.ABL camp.ABL two.ACC there legions.ACC left.3S.PRF and
partem *auxiliorum*
part.ACC auxiliaries.GEN
'With the camp fortified, he left two legions and a part of the auxiliaries there.' (Caes. Gal. 1.49.4)

Finally, since gap degree 2 examples arise naturally, even without clitics, there are examples where a clitic intrudes in an otherwise degree 2 discontinuity, yielding gap degree 3. And there are a few gap 3 examples without clitics. We refrain from showing examples here, as they inevitably get quite complex.

When it comes to illnestedness, we saw in the previous section that examples are extremely rare in the PROIEL treebank. Nevertheless, it is worth pointing out that the ones that do occur look perfectly 'natural', in the sense that it is hard to come up with alternative analyses that make linguistic sense and capture the sentence structure without an illnested dependency. (12) shows an example where the subject appears inside the object NP, at the same time as there is an extraposed relative clause belonging to it.

(12) *Magnam* *Caesarem* *iniuriam* *facere* *qui* *suo adventu* *…*
great.ACC Caesar.ACC injustice.ACC do who.NOM by his arrival
'(He said that) Caesar was doing a great injustice, who by his arrival …' (Caes. Gal. 1.36.4)

Taken together with the metrical data discussed in section 3.1, this suggests that illnestedness is not ungrammatical in Latin, although it clearly is strongly dispreferred (in prose).

4 So how complex is Latin really?

We can clearly conclude that Latin is not a tree-adjoining language. As Table 4 shows, there are simply too many trees of gap degree > 1, and examples such as (8)–(11) show that these arise through combinations of well-established processes of Latin grammar.

However, Table 4 also shows that we do not really need the ability of LFG to transport unbounded amounts of features across unbounded distances in the tree to capture the data found in Latin treebanks: Trees of gap degree 3 are already very rare. Nevertheless they arise through well-defined grammatical processes (and are not, for example, artefacts of the annotation scheme). It is therefore impossible to define a theoretical upper bound on gap degree in Latin.

In a way, the situation is analogous to what we see in center-embedded recursion. We can deal with finite levels of center-embedded recursion in a regular (finite state) grammar by adding states to the automaton. And center-embedding of more than three levels turns out to be nonexistent in corpora (Karlsson 2007), so for practical purposes, the finite state approach could work. But linguists prefer context-free grammars both because it is hard or impossible to define a theoretical upper bound on the levels of center-embedding and, crucially, because analyses cast in terms of a CFG are linguistically more perspicuous. A similar argument applies, I contend, to syntactic discontinuities: although we could deal with them in practical terms – at least when we confine the attention to the texts in the existing Latin treebanks – by adopting a k-MCFG as our formalism for some (quite small) k, it is hard to argue theoretically for any particular k and – as we have already seen – analyses that are cast in more expressive formalisms can turn out to be more intuitive. In other words, we can adapt Harris' argument for assuming infinite levels of center embeddings to unbounded discontinuous dependencies: fixing a k is a "highly arbitrary and numerical condition" that has no place in linguistic theory. In that respect, 2 – the number that (restricted to wellnested dependencies) would give us the expressive power of TAG or classical CCG – is no different from any other number.

This gives us an argument for adopting LFG as a formalism even if that is expressive overkill in practical terms.[9] And although LFG does not provide an obvious way of restricting discontinuities, we will see that it does provide a way of analyzing them that gives us a natural metric for discontinuity complexity in the form of the number of reentrancies they require. Consider first the mock Latin sentence in (13).

(13) *Maximilianus trusit bonum Fredericum*
Max.NOM pushed good.ACC Fredrick
Maximilian pushed good Fredrik

In LFG terms, this can be analyzed with the c- and f-structure in Figure 5. A characteristic feature of this is that the ϕ mapping from maximal projections (S and NP) is injective: there are no reentrancies, i.e. distinct maximal projections mapping to the same f-structure.

9 There may be an intermediate formalism available: As shown by Kallmeyer and Satta (2009), Tree-Tuple Multiple Component TAGs (TT-MCTAGs) can describe German scrambling and have a polynomial parsing algorithm.

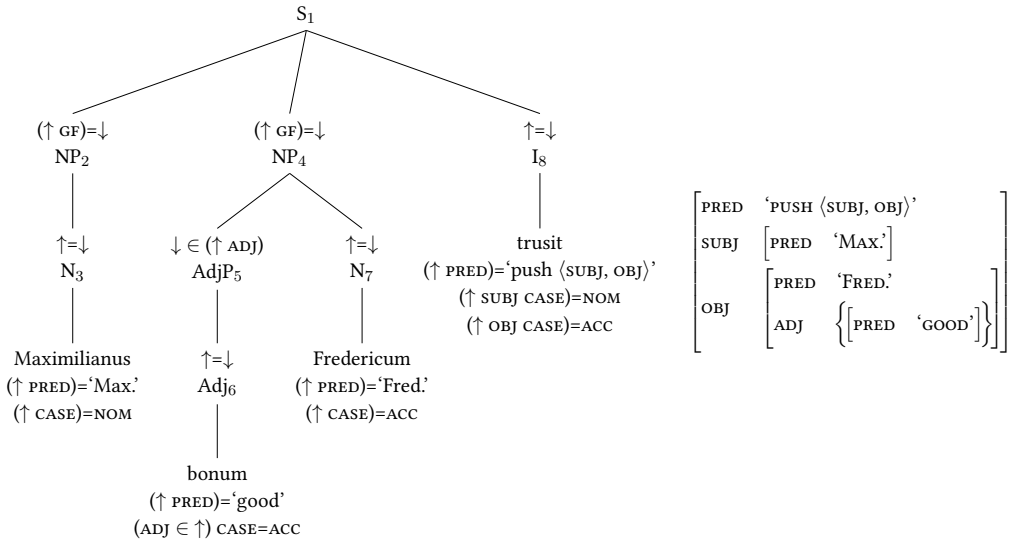


Figure 5: C- and f-structure for (13)

Now consider what happens if we permute *trusit* and *bonum* to yield a discontinuous c-structure. If we want to avoid non-local assignment of grammatical functions, the obvious way to achieve this is by using a c-structure embedding as in Figure 6. This introduces a reentrancy: for this c-structure to yield the correct f-structure (namely the same as in Figure 5), we must make sure that NP_4 and NP_8 map to the same f-structure, i.e. GF on both these nodes must be resolved to the same grammatical function.[10] In other words, the syntactic discontinuity is mirrored by structural complexity in the form of a reentrancy. Obviously, if we had yet another discontinuous dependent of *Fredericum* that was discontinuous from *bonum* so that we had a gap degree 2 discontinuity, we would need another reentrancy to capture that.

Now observe what happens if we have a deeper gap as in (14). This yields the c-structure in Figure 7. We observe that the extra depth of the discontinuity yields an extra reentrancy as compared with the otherwise similar discontinuity in Figure 6, for to get the correct f-structure from Figure 7, we must map both NP_4 and NP_5 to the same f-structures as NP_9 and NP_{10} respectively.

(14) *Maximilianus* *boni* *trusit* *Frederici* *filium*
Max.NOM good.GEN pushed Fredrick.GEN son.ACC
Maximilian pushed good Fredrick's son

10 In this particular example, case agreement will ensure that.

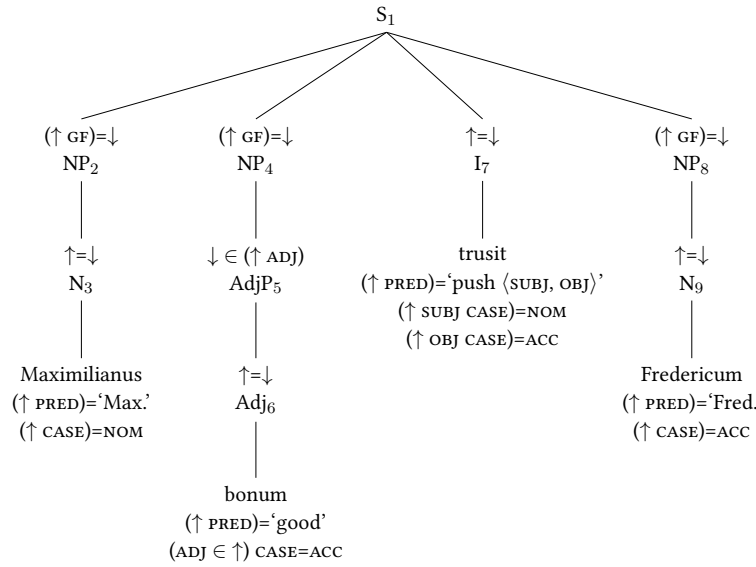


Figure 6: C-structure for permuted version of (13)

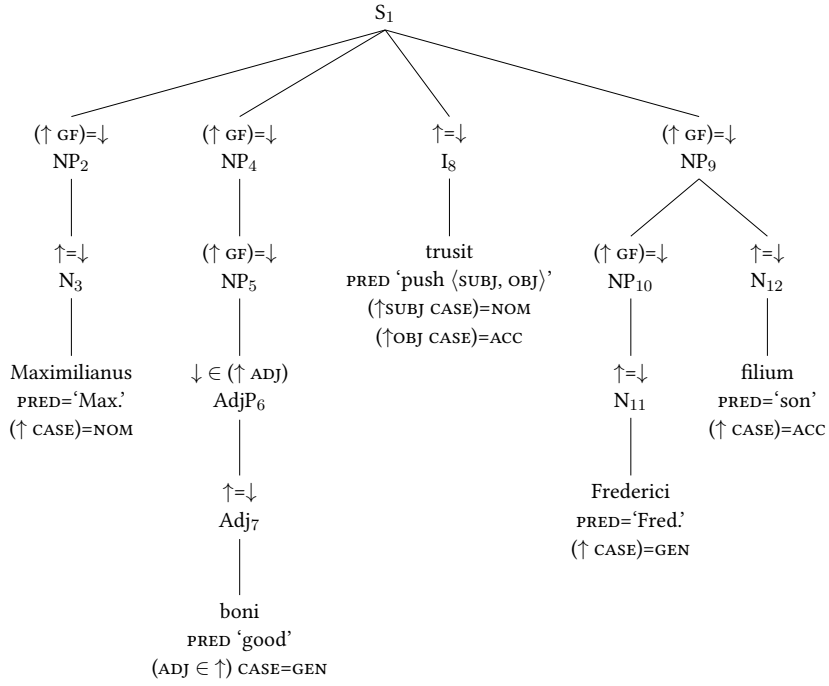


Figure 7: C-structure for (14)

The c-structure in Figure 7 clearly violates the principle in (6). Nevertheless, it is an attractive analysis compared with, say, an analysis where AdjP_6 would be directly embedded under S and annotated with $(\uparrow \text{GF}^+)$ because that would introduce non-local constraints and because Figure 7 wears its complexity on its sleeves, in the form of the length of the unit branch under NP_4. Like the analysis of Dutch cross-serial dependencies, this relies on zipper unification. As pointed out by Maxwell and Kaplan (1996, p. 24), zippers introduce computational complexity because they mean that depth of f-structures that must be unified can grow as a function of the length of the sentence. (In fact, because we allow a cyclic unit branch, it can grow even without the sentence increasing in length.) But this is really a practical problem and as such it allows a practical solution, namely a brute force bound on the length of zippers. And that is where the treebank data become interesting, for they suggest that this bound can be set quite low.

5 Conclusion and challenge

In sum, we have seen that extant Latin treebanks display syntactic discontinuities that require us to go beyond the capacity of well-known mildly context-sensitive grammar formalisms such as CCG and TAG. It has already been argued on theoretical grounds that these formalisms cannot capture data such as German scrambling (Becker et al. 1992). But as pointed out by Kuhlmann (2013), formalisms (weakly) equivalent to TAG still have very good coverage on treebanks. That, however, is not the case in Latin (and still less so in Ancient Greek), thereby verifying the inadequacy of TAG on actual treebank data.

From a theoretical point of view, this means that trees of gap degree 1 have no particular theoretical importance. Rather, the corpus data suggests that gap degrees (and depths) have a Zipfian distribution that quickly decreases beyond 1. So there is no theoretical reason to stay with k-MCFGs. And in fact we have seen that although LFG parsing is intractable, the formalism reflects the complexity of syntactic discontinuities in a rather nice and intuitive way, paving the way for empirical studies on how much of the theoretically desired expressivity is actually needed for practical purposes.

One challenge remains: we have seen that illnested discontinuities are strongly dispreferred in most treebanks, with an exception for Latin poetry. But unlike gap degree and depth, illnestedness does not correspond to any complexity in the LFG formalism. In other words, LFG as it currently stands lacks the theoretical resources to express the strong dispreference that we observe in corpora.

References

Bamman, David and Gregory Crane (2011). "The ancient Greek and Latin dependency treebanks". In: *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*. Berlin: Springer Verlag, pp. 79–98.

Becker, Tilman, Owen Rambow, and Michael Niv (1992). *The derivational generative power of formal systems or scrambling is beyond LCFRS*. IRCS Report 92-38. Institute for Research in Cognitive Science, University of Pennsylvania.

Bresnan, Joan (2001). *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics 16. Oxford: Blackwell.

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler (2015). *Lexical-Functional Syntax*. Malden: John Wiley & Sons.

Bresnan, Joan, Ronald M. Kaplan, Stanley Peters, and Annie Zaenen (1982). "Cross-serial dependencies in Dutch". In: *Linguistic Inquiry* 13, pp. 613–635.

Chomsky, Noam (1956). "Three models for the description of language". In: *IRE Transactions on Information Theory* 2.3, pp. 113–124.

Clark, Alexander (2014). "An introduction to multiple context free grammars for linguists". http://www.cs.rhul.ac.uk/home/alexc/lot2012/mcfgsforlinguists.pdf.

Culy, Christopher (1985). "The complexity of the vocabulary of Bambara". In: *Linguistics and Philosophy* 8.3, pp. 345–351. DOI: 10.1007/BF00630918.

Dyvik, Helge (1968). "Pauli gallinae". Available from http://folk.uib.no/hfohd/SLF/Dyvik/gallinae/gallinae.html.

Francez, Nissim and Shuly Wintner (2012). *Unification Grammars*. Cambridge: Cambridge University Press.

Gaifman, Haim (1965). "Dependency systems and phrase-structure systems". In: *Information and Control* 8.3, pp. 304–337.

Gazdar, Gerald (1981). "Unbounded dependencies and coordinate structure". In: *The Formal Complexity of Natural Language*. Ed. by W.J. Savitch, E. Bach, W.E. Marsh, and G. Safran-Naveh. Springer, pp. 183–226.

Gibson, Edward (2000). "The dependency locality theory: A distance-based theory of linguistic complexity". In: *Image, Language, Brain*. Ed. by Alec P. Marantz, Yasushi Miyashita, and Wayne O'Neil. Cambridge, MA: MIT Press, pp. 95–126.

Goldstein, David M. and Dag T. T. Haug (2016). "Second-position clitics and the syntax-phonology interface: The case of ancient Greek". In: *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar, Polish Academy of Sciences, Warsaw, Poland*. Ed. by Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King, and Stefan Müller. Stanford, CA: CSLI Publications, pp. 297–317.

Harris, Zellig S. (1957). "Co-occurrence and transformation in linguistic structure". In: *Language* 33.3, pp. 283–340.

Haug, Dag T. T. (2015). "Treebanks in historical linguistic research". In: *Perspectives on Historical Syntax*. Ed. by Carlotta Viti. Amsterdam: Benjamins, pp. 188–202.

Haug, Dag T. T. and Marius L. Jøhndal (2008). "Creating a parallel treebank of the old Indo-European bible translations". In: *Language Resources and Evaluation*. Marrakech, Morocco, pp. 27–34.

Havelka, Jiří (2007). "Beyond projectivity: multilingual evaluation of constraints and measures on non-projective structures". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 608–615.

Johnson, Mark (1986). "The LFG treatment of discontinuity and the double infinitive construction in Dutch". In: *Proceedings of the Fifth West Coast Conference on Formal Linguistics*. Ed. by Mary Dalrymple, Jeffrey Goldberg, Kristin Hanson, Michael Inman, Chris Piñon, and Stephen Wechsler. Stanford Linguistics Association. Stanford, pp. 102–118.

Kallmeyer, Laura and Giorgio Satta (2009). "A polynomial-time parsing algorithm for TT-MCTAG". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*. ACL '09. Suntec, Singapore: Association for Computational Linguistics, pp. 994–1002.

Kaplan, Ronald M. and Joan Bresnan (1982). "Lexical-Functional Grammar. A formal system for grammatical representations". In: *The Mental Representation of Grammatical Relations*. Ed. by Joan Bresnan. Cambridge, MA: MIT Press, pp. 173–281.

Karlsson, Fred (2007). "Constraints on multiple center-embedding of clauses". In: *Journal of Linguistics* 43.02, pp. 365–392.

Kuhlmann, Marco (2013). "Mildly non-projective dependency grammar". In: *Computational Linguistics* 39.2, pp. 355–387.

Kuhlmann, Marco, Alexander Koller, and Giorgio Satta (2015). "Lexicalization and generative power in CCG". In: *Computational Linguistics* 41.2, pp. 215–247. DOI: 10.1162/COLI_a_00219.

Kuhlmann, Marco and Joakim Nivre (2006). "Mildly non-projective dependency structures". In: *Proceedings of the COLING/ACL Main Conference Poster Sessions*. COLING-ACL '06. Sydney, Australia: Association for Computational Linguistics, pp. 507–514.

Maier, Wolfgang and Timm Lichte (2011). "Characterizing discontinuity in constituent treebanks". In: *Formal Grammar: 14th International Conference, FG 2009, Bordeaux, France, July 25-26, 2009, Revised Selected Papers*. Ed. by Philippe de Groote, Markus Egg, and Laura Kallmeyer. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 167–182. DOI: 10.1007/978-3-642-20169-1_11.

Mambrini, Francesco and Marco Passarotti (2013). "Non-projectivity in the Ancient Greek dependency treebank". In: *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*. Prague, Czech Republic: Charles University in Prague, MatfyzPress, Prague, Czech Republic, pp. 177–186.

Martens, Scott and Marco Passarotti (2014). "Thomas Aquinas in the TüNDRA: Integrating the Index Thomisticus treebank into CLARIN-D". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn

Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA).

Maxwell, John T. and Ronald M. Kaplan (1996). "Unification-based parsers that automatically take advantage of context freeness". In: *Proceedings of the LFG'96 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications, pp. 1–31.

Pullum, Geoffrey (1982). "Free word order and phrase structure rules". In: *Proceedings of the Twelfth Annual Meeting of the North Eastern Linguistic Society*. Ed. by James Pustejovsky and Peter Sells. Graduate Linguistics Students Association. Amherst, MA, pp. 209–220.

– (1986). "Footloose and context-free". In: *Natural Language and Linguistic Theory* 4, pp. 409–414.

Ross, John R. (1967/1986). *Infinite Syntax*. Reprint of 1967 MIT dissertation. Norwood, NJ: Ablex.

Shieber, Stuart M. (1985). "Evidence against the context-freeness of natural language". In: *Linguistics and Philosophy* 8, pp. 333–343.

Zaenen, Annie and Ronald M. Kaplan (1995). "Formal devices for linguistic generalizations: West Germanic word order in LFG". In: *Formal Issues in Lexical-Functional Grammar*. Ed. by Mary Dalrymple, Ronald M. Kaplan, and John T. Maxwell. Stanford: CSLI Publications, pp. 215–239.

Increasing grammar coverage through fine-grained lexical distinctions

Petter Haugereid

Abstract. In this paper, I show how the development of Norsyg, an HPSG-inspired constructionalist grammar of Norwegian, benefits from the highly specific and precise implementation of NorGram, an LFG grammar for Norwegian. I focus on one aspect, NorGram's fine-grained lexical categories. The aim of the paper is twofold: (i) to give a glimpse of the process of developing a computational grammar, and (ii) to illustrate how a constructionalist grammar benefits from the insights of NorGram, even though the grammatical models differ significantly.

1 Introduction

There are different approaches to the automatic analysis of sentences. A common approach is to train a shallow statistical parser based on large amounts of syntactically annotated texts (a treebank). The advantage of these systems is that they utilize already existing resources (annotated texts), they have large coverage, and they are very fast. The main problems are that once they have reached a certain level, it is hard to make improvements, and there is never a guarantee that the analysis provided is the correct one, or even a possible one.

A different approach to automatic analysis of sentences is to develop deep, rule-based computational grammars. These systems take much more time to develop, and in the beginning, the coverage is very limited. If there is a missing lexical item or a missing rule for a certain linguistic construction, the grammar does not provide a parse. Scaling up these systems may take several years. However, given the fact that the systems are rule based, the grammar developer is in control of what analyses are possible, and corrections that address particular linguistic phenomena may be made. So if the aim of the system is high precision, building a deep grammar is a better option than building a shallow parser in the long term.

One possible reason for the relatively high interest in linguistically founded computational grammars in Norway is that treebanks required for building statistical parsers

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1334.

have not been available for Norwegian until recently.[1] Another reason is the interest in grammar formalisms like LFG and HPSG, which are both associated with environments for grammar implementations. In this paper I will describe two quite different computational grammars, NorGram and Norsyg, and show how insights in NorGram can be used to develop the coverage of Norsyg.

2 NorGram and Norsyg

NorGram (Dyvik 2000) is the result of a long-term, incremental effort to develop a theoretically motivated, large coverage grammar for Norwegian.[2] It is written within the framework of Lexical Functional Grammar (LFG) (Bresnan 2001), under the ParGram umbrella (Parallel Grammar Project), which is an association of groups working on computational LFG grammars for various languages (Butt et al. 2002). LFG is a lexicalist framework where linguistic objects are represented with mainly two structures: c-structure (constituent structure) and f-structure (functional structure). The c-structure shows the hierarchical organization of constituents in a clause at the same time as it shows how the parser has worked, combining constituents by means of phrase structure rules. The f-structure represents linguistic information about each constituent and shows the functional relationship between the constituents. In this paper we will mainly consider c-structures. The tree in Figure 1 shows the c-structure of the main clause in example (1) analyzed with NorGram.[3]

(1) *Dessverre ville ikke disse studentene lære syntaks.*
unfortunately wanted not these the students learn syntax
'Unfortunately, these students didn't want to study syntax.'

Norsyg is a typed feature structure grammar, and is implemented with the LKB system (Copestake 2002) as a part of the DELPH-IN effort (http://www.delph-in.net/). It is based on the Grammar Matrix (Bender et al. 2002), which is a starter kit for HPSG grammar development. Norsyg has kept most of the HPSG feature geometry, but the intuition behind the analyses is radically different from regular lexicalist HPSG grammars. It is a constructionalist grammar, and the backbone of the grammar consists of about 15,000 constructions. Argument-frame constructions constitute the main part of

1 NorGramBank (Dyvik et al. 2016), a large-scale syntactically annotated treebank of Norwegian, has recently become available. NorGramBank has been developed through the projects TrePil (Rosén, De Smedt, Dyvik, et al. 2005) and INESS (Rosén, De Smedt, Meurer, et al. 2012). This treebank is based on parses obtained with NorGram, which will be further discussed in the present paper.

2 Helge Dyvik has been a pioneer of computational grammar and parsing in Norway. He started already in the 1980's with the D-PATR formalism, a development environment for unification-based grammars. This work was carried over in the PONS project (Dyvik 1989), a machine translation project with semantic transfer. Since 1999, he has been the main developer of NorGram, which was used as a parsing grammar not only for NorGramBank but in the translation projects LOGON (Oepen et al. 2007) and HandOn.

3 The tree is slightly simplified for expository reasons.

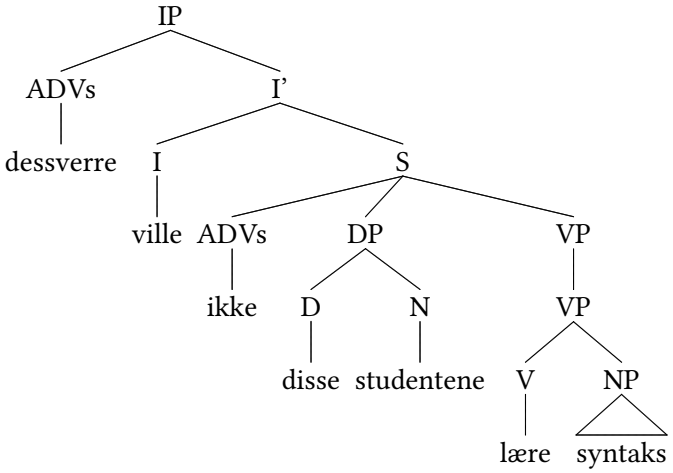


Figure 1: LFG c-structure of a Norwegian main clause with NorGram, cf. example (1)

these; each argument frame of each verb is assumed to be a construction. For example, the transitive and ditransitive frames of the verb *lære* 'learn' are assumed to be unique constructions.

The high number of constructions, in addition to well known phenomena such as flexibility with regard to positioning of adjuncts, the active–passive voice alternation, and different kinds of clause structures, make the assumption of flat phrase structure rules impossible. Instead, the grammar is given a "fragmented" design, where constructions are assumed to be built up of *subconstructions*. A subconstruction may be a binary phrase structure rule with a word as its second daughter (see the rules in Figure 2), or it may be a lexical item, like a verb, a function word or an idiomatic word. Each subconstruction contributes a simple type, which by itself may carry very little meaning. During parsing, however, the types provided by the subconstructions are unified, and if the parse succeeds, the unification of the subconstruction types yields one of the 15,000 construction types licensed by the grammar.[4] This subconstructional design gives the grammar the flexibility needed to accommodate a wide variety of syntactic phenomena while limiting the number of phrase structure rules to 110. The tree in Figure 2 shows the parse tree of the main clause in (1) analyzed with Norsyg.

The trees in Figure 1 and Figure 2 illustrate a principal difference between the two grammars. NorGram on the one hand is based on standard X-bar theory. It consists of phrase structure rules such as VP → V NP. The grammar relies heavily on the formula-

4 Needless to say, the type hierarchy where the possible combinations of subconstruction types are defined, is rather big, but once it is compiled, its size does not affect the efficiency of the parser very much. A small test indicates an increase in parsing time of about 15% when the lexicon is scaled up.

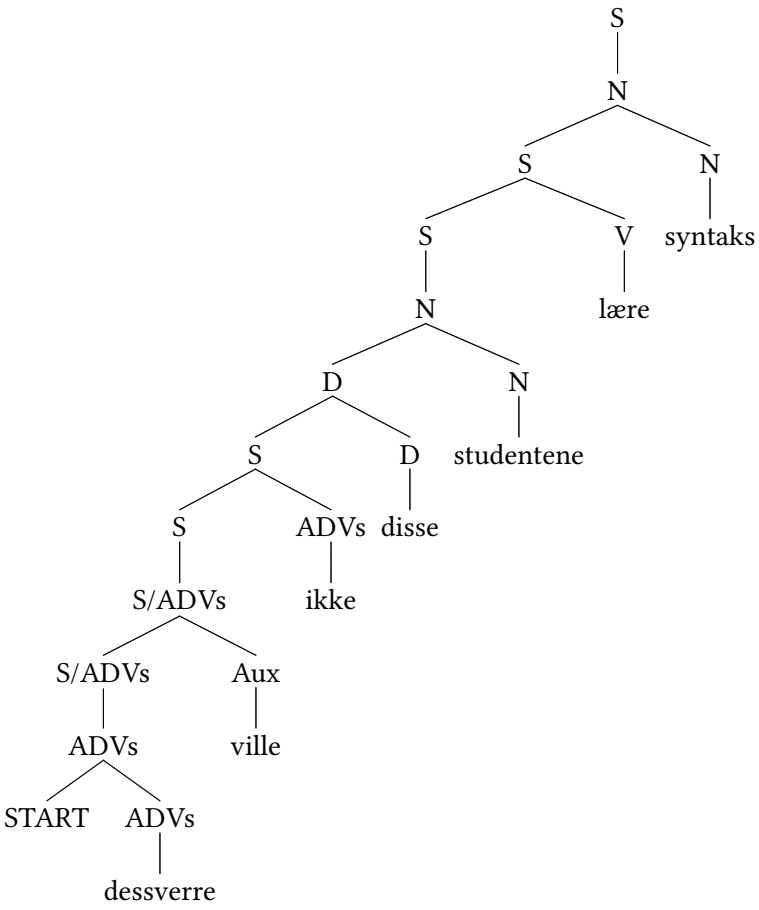


Figure 2: Parse tree of a Norwegian main clause with Norsyg, cf. example (1)

tion of this kind of phrase structure rules. The structures are relatively flat, a rule may have more than five daughters, and phrasal constituents may appear in a non-initial and non-final position, like the DP *disse studentene* in Figure 1.

Norsyg on the other hand consists of only binary and unary phrase structure rules where the second daughter (if there is one) is a word. As the words combine, a feature structure is built where linguistic information of the clause is represented. From the resulting feature structure, the constituent structure in Figure 3 is derived.[5] The parse also results in a semantic representation, an MRS (Copestake et al. 2005).

5 By separating the parse tree (see Figure 2) from the constituent tree (see Figure 3), Norsyg allows for flat constituent structures at the same time as the phrase structure rules are unary or binary. The motivation behind this separation is explained in Haugereid and Morey (2012).

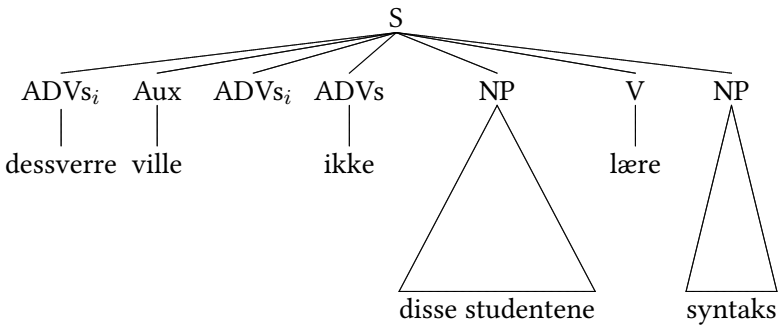


Figure 3: Constituent structure of a Norwegian main clause by Norsyg

Allthough Norsyg has been developed over many years, there are several phenomena that are covered by NorGram that are yet to be implemented in Norsyg. In the next section I will show how I use NorGramBank, the treebank syntactically annotated with the help of NorGram, to identify phenomena that will have an impact on the coverage of Norsyg.

3 Identifying phenomena not covered by Norsyg

In grammar development, identifying phenomena that are not covered by the grammar, is usually very easy. One can just take a random sentence that the grammar does not parse, and use the diagnostic tools of the parser to find out what goes wrong. This is from my own experience the most common way to improve coverage of a grammar, and is the approach one would take during treebanking. If a sentence does not parse, one attempts to make it parse. Often adding a lexical item is enough, but one may also encounter very challenging phenomena which require a complete overhaul of the grammar. However, the impact of the changes made, whether they are large or small, may be just a very small increase in coverage.

In this section, I describe a more systematic way of identifying phenomena that are not covered by Norsyg, and which will have an impact on the coverage of the grammar. And in this, I will utilize the linguistic insights behind NorGram and the 60 million word corpus NorGramBank which has been syntactically annotated with the help of NorGram.

Syntactic rules in LFG have information about the c-structure of its constituents as well as the f-structure. In the parsing process, first the c-structure backbone is constructed (as a packed parse forest), and then, in the next step, the (f-structure) equations attached to the c-structure rules are solved. In order to reduce the size of the c-structure parse forest, NorGram has been equipped with a relatively large set of dis-

criminative preterminals or lexical categories. This results in fewer equations that have to be solved, and makes the parser more efficient.

In order to see which preterminals are used by the grammar, I downloaded a frequency list of preterminals from NorGramBank. It turns out that the treebank has 220 different preterminals. A few of them have a grammar internal function (different kinds of tags for enhancing processing), and others represent punctuation marks. However, most of the preterminals have a solid linguistic foundation. The most frequent lexical categories are given in Table 1.[6] We can see that the most frequent categories are nouns (12.82%) and pronouns (8.83%). The table also shows that the lexical categories are fine-grained. There are for example separate categories for finite main verbs (Vfin), finite auxiliaries (Vauxfin), and finite copula verbs (Vcopfin).

Nr.	Category	Freq.
1	N	12.82
2	PRON	8.83
3	Vfin	7.28
5	P	6.53
6	A	5.84
8	Vauxfin	3.14
9	Vinf	2.95
10	Vcopfin	2.16

Table 1: The eight most frequent lexical categories in NorGramBank, with frequencies in percentages

Further down the list, there are some lexical categories that represent phenomena that are not covered or treated in a systematic way by Norsyg. The categories are well documented in the NorGram online documentation, and I use this documentation and my knowledge of Norsyg to identify the missing categories. The most frequent are shown in Table 2: finite inquit verbs (Vinqfin), prepositions that take subordinate clauses or infinitival clauses as complements (Pvbobj), interjections (INTERJ), correlative coordinators (CONJcorr), and titles (TTL).

In the following, I will show how I have utilized the information about the categories that are unaccounted for in the development of Norsyg. Given the constructionalist design of Norsyg, some of the phenomena will not be analyzed in terms of separate lexical categories, like inquit verbs and prepositions that take subordinate clauses or infinitival clauses as complements. Rather, the phenomena will be accomodated by constructions. Prepositions that take subordinate clauses or infinitival clauses as complements, for example, will still have the lexical category preposition, but they will

6 The fourth and seventh most frequent preterminals are PERIOD and COMMA. They are left out of the table.

Nr.	Category	Freq.
41	Vinqfin	0.40
42	Pvbobj	0.29
67	INTERJ	0.08
68	CONJcorr	0.07
72	TTL	0.06

Table 2: The five most frequent NorGram lexical categories unaccounted for in Norsyg

be made compatible with constructions that involve a subordinate clause or infinitival clause complement.

4 Developing Norsyg on the basis of NorGram lexical categories

As mentioned, NorGramBank is a corpus of approximately 60 million words. The larger part of the corpus has been stochastically disambiguated, and approximately 315,000 words of parsed text are manually disambiguated. From the corpus of manually disambiguated sentences, I have selected 14,770 sentences marked as "gold" by the annotators (127,644 words). These are sentences that are parsed by NorGram and that have been disambiguated by an annotator and marked as correct.

In my work on adding missing analyses to Norsyg I started with the lexical categories on top of the list in Table 2, inquit verbs, and worked my way down. For each phenomenon I added to the grammar, I did a test run on the gold corpus to check the impact of the changes made. Before the development started, I checked Norsyg's coverage of the gold corpus. It parsed 7216 of the 14770 sentences (48.86%).

4.1 Inquit verbs (Vinq)

Inquit verbs are a group of verbs that typically indicate direct speech, as shown in (2), but this group also includes verbs with the same syntactic behavior, like *tro* 'believe' and *synes* 'think'.

(2) *Jeg sov, sa han.*
I slept said he
'I slept, he said.'

This is a phenomenon that had not been implemented in Norsyg, and in order to account for the construction, three rules were introduced. One rule is created for sentences where the inquit complement is a full sentence, as in (2). There is also a rule where the complement is some other constituent, such as an NP or PP, or an interjection. The third rule marks the position where the complement is extracted from. In

addition, the 107 verbs marked as regular inquit verbs in Norgram were constrained in such a way that they were allowed as verbs in inquit constructions in Norsyg.

After adding the new rules and lexical constraints to the grammar, 100 sentences that earlier did not get a parse, now were parsed by Norsyg, an increase of 0.67%. Some examples are given in (3)–(5).

(3) *Nå er jeg trygg, sier hun og smiler mot sykepleieren.*
now am I safe says she and smiles towards the nurse
'Now I am safe, she says and smiles towards the nurse.'

(4) *De liker meg ikke, skriver hun.*
you like me not writes she
'You don't like me, she writes.'

(5) *En syk øvelse, tenker hun sint.*
a sick exercise thinks she angry
'A sick exercise, she thinks angrily.'

An abbreviated Norsyg analysis of (4) is provided in Figure 4. It shows the application of two of the added rules. (The mother nodes of the rules are framed.) The top rule in the tree is the rule that marks the position the complement is extracted from. It is a unary rule that takes as input a structure which has a sentence on the SLASH list.[7] The framed S/S rule further down the tree is a rule that takes as input a main clause and a comma, and enters selected features of the main clause onto the SLASH list.

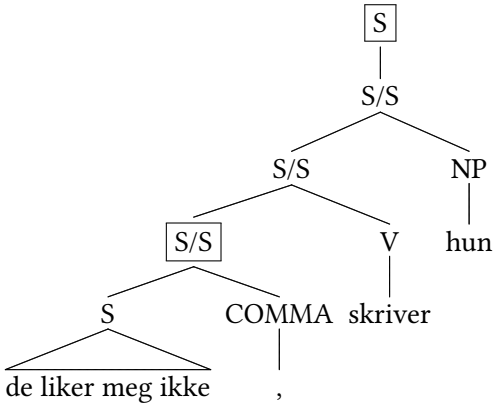


Figure 4: Norsyg analysis of sentence with inquit verb and main clause complement

7 In HPSG, long distance dependencies are handled by a feature SLASH. Contrary to regular HPSG grammars, the extraction site in Norsyg dominates the topic, rather than the other way round. This is enforced by the incremental bottom-up parsing process.

4.2 Pvbobj: prepositions that take subordinate clauses or infinitival clauses as complements

In the second batch, I added analyses for Pvbobj, the category for prepositions that take subordinate clauses or infinitival clauses as complements to form PPs that function as adverbials. There are 23 such prepositions, among them *for* 'for', *i tillegg til* 'in addition to' and *uten* 'without'. The inclusion of a new analysis for these prepositions in Norsyg involved changing the constraints of these prepositions so that they were allowed in constructions where the head is a preposition and the complement is a subordinate clause or an infinitival clause.

After the analyses were added, the grammar produced analyses for 49 more sentences, among them, the sentence in (6).

(6) *Amygdala starter analyser for å se mulige farer.*
Amygdala initiates analyses for to see possible dangers
'Amygdala initiates analyses in order to see possible dangers.'

An abbreviated analysis of (6) is presented in Figure 5. The preposition with the new constraints is framed.

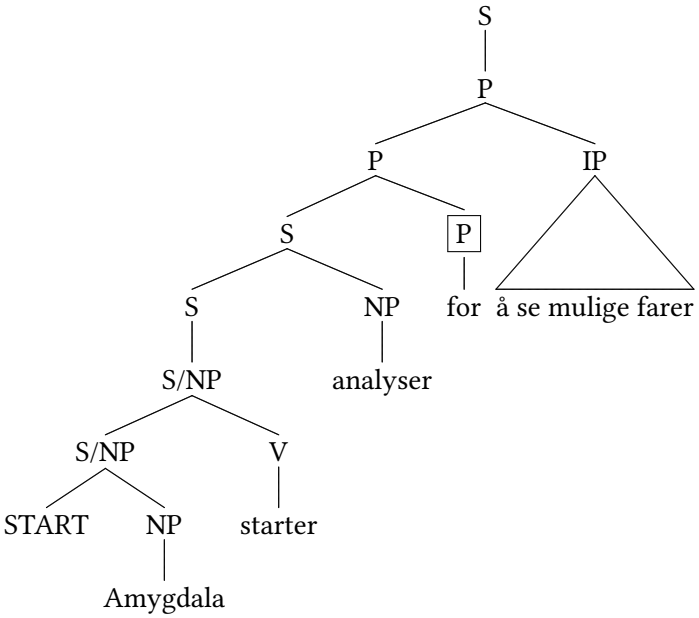


Figure 5: Norsyg analysis of sentence with a PP adjunct that has an infinitival clause complement

4.3 INTERJ: Interjections

In the third batch, I added analyses for interjections. In Norsyg, they are given the status of roots, which means that they are allowed to form sentences on their own, function as arguments of inquit verbs, or be coordinated with other roots (including sentences).

After the analyses were added, the grammar produced analyses for 31 more sentences, among them, the sentences in (7) and (8). The sentence in (8) also benefits from the recently added inquit analysis.

(7) – *Å, har du ikke hørt det?*
oh have you not heard it
'– Oh, haven't you heard?'

(8) – *Jada, mamma, fleipet han.*
of course mum he joked
'– Of course, mum, he joked.'

An abbreviated analysis of (7) is provided in Figure 6. It shows the new category of interjections (INTERJ) and the new rule for adding interjections (S). The rule that adds interjections, takes a START symbol as its first daughter and an interjection as its second daughter and forms a structure with root status. It is then coordinated with the following yes–no question.[8]

4.4 CONJcorr: correlative coordinators

In the fourth batch, the words *både* 'both', *verken* 'neither', and *såvel* 'both' were given an analysis. These are words that initiate a coordination and select the coordinator between the conjuncts. There had been an analysis for these words at an earlier stage, but it had become obsolete. After recreating the analysis, 11 more sentences got a parse, among them (9) and (10).

(9) *Den knuste både negl og bein.*
it broke both nail and bone
'It broke both nail and bone.'

(10) *Både han og jeg har fått tørre føtter.*
both he and I have got dry feet
'Both he and I have got dry feet.'

8 Given the left-branching design of the grammar, the parse tree of the coordinated structures looks rather counter-intuitive, but it is chosen in order to maintain the overall incremental design. It also makes possible a novel account of gapping constructions (Haugereid 2017).

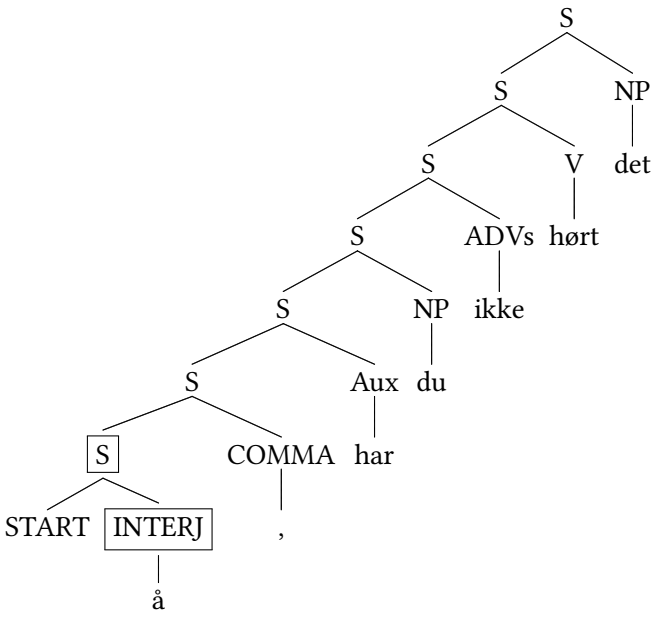


Figure 6: Norsyg analysis of sentence with interjection coordinated with yes–no clause

4.5 TTL: title

In the fifth batch, 5633 titles from Norgram were added, *dr.* 'dr.', *språkprofessor* 'linguistics professor', *fagekspert* 'professional', *norrønfilolog* 'Old Norse philologist', *hedersmann* 'man of honour', *ektemann* 'husband', *bestefar* 'grandfather', *onkel* 'uncle', and *pappa* 'daddy', among others.[9] They are analyzed as pre-modifiers of proper nouns in Norsyg. After adding the analysis of the titles, 24 new sentences were parsed by the grammar, among them the sentences in (11) and (12).

(11) *Mormor snudde seg og så mot onkel Ernst.*
Grandma turned and looked so towards uncle Ernst
'Grandma turned and looked in the direction of uncle Ernst.'

(12) – *Jomfru Bendeke, mener du?*
virgin Bendeke, mean you
'– Miss Bendeke, you mean?'

9 As it happens, all these titles can be attributed to Helge Dyvik.

5 Results

The effect of the grammar development work is summarized in Table 5. It shows an increase of coverage on the gold corpus of 210 sentences, or 1.41%. It also shows that the categories at the top of the list resulted in the most significant gains of coverage.

phenomenon	coverage	%
Before	7216	48.86
Inquit verbs	7316	49.53
Pvbobj	7365	49.86
INTERJ	7396	50.07
CONJcorr	7407	50.14
TTL (title)	7426	50.27

The effect is also illustrated in the chart in Figure 5. It shows that the number of new sentences that receive an analysis by the grammar increases as the frequency of the added lexical category goes up. When the frequency is 0.06 % (titles), the number of added sentences is 24, and when the frequency is 0.4 % (inquit verbs), the number of new analyses is 100. This is of course an expected result, and it confirms the obvious, namely that there is more to gain from adding analyses for more frequent lexical categories.

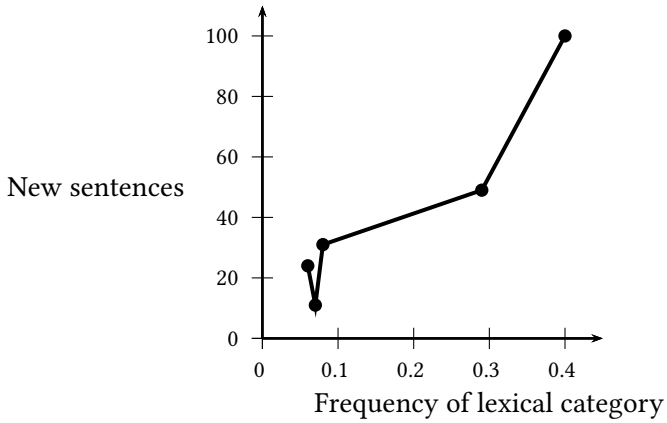


Figure 7: Number of new sentences with regard to frequency of added lexical category

A manual inspection of the sentences that earlier did not get a parse, but which after the changes to the grammar got a parse, shows that about two out of three sentences can be attributed directly to the added analysis. The last third is mainly made up of longer sentences for which the chart size after the changes to the grammar no longer

exceeds 20 megabytes.[10] There is also a set of sentences which after the changes to the grammar exceeds the 20 megabyte limit, and these sentences therefore do not get an analysis. The measures are therefore not completely precise, but still they give a clear indication of the impact of the added analyses.

It should also be mentioned that a number of sentences which earlier were given an analysis for the wrong reason, got the correct analysis after the changes to the grammar. These changes are however difficult to measure.

6 Conclusion

Using the lexical categories of NorGram in the development of Norsyg has proved to be a fruitful exercise. As a grammar writer, I can foresee what grammatical phenomena the grammar I am developing can account for. However, it is harder to see exactly which changes will amount to the greatest gain of coverage. Here, the well-documented lexical categories of NorGram have been very useful. By looking at their frequencies in the NorGramBank corpus, and consulting the detailed online documentation and the analyses provided by NorGram, I have been able to pick five categories that represent phenomena not covered by the grammar and increase its coverage by 1.41%.

Acknowledgments

I would like to thank the research group Språk og samfunn at Western Norway University of Applied Sciences, two anonymous reviewers, and Victoria Rosén for very useful comments and suggestions for this paper. I would also like to thank Helge Dyvik for being a great inspiration and for his contribution to the field of computational linguistics.

References

Bender, Emily M., Dan Flickinger, and Stephan Oepen (2002). "The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars". In: *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*. Ed. by John Carroll, Nelleke Oostdijk, and Richard Sutcliffe. Taipei, Taiwan, pp. 8–14.

Bresnan, Joan (2001). *Lexical-Functional Syntax*. Malden, MA: Blackwell.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer (2002). "The Parallel Grammar Project". In: *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Com-*

10 The maximum chart size is set to 20 megabytes during batch parsing of the 14770 'gold' sentences in order to limit the batch parsing time to about 30 minutes. By setting the chart size to 200, the number of sentences parsed is increased by 479, but this also increases the batch parsing time to almost three hours.

putational Linguistics (COLING), Taipei, Taiwan. Ed. by John Carroll, Nelleke Oostdijk, and Richard Sutcliffe. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1–7.

Copestake, Ann (2002). *Implementing Typed Feature Structure Grammars*. Stanford, CA: CSLI publications.

Copestake, Ann, Dan Flickinger, Carl J. Pollard, and Ivan A. Sag (2005). "Minimal Recursion Semantics: an Introduction". In: *Research on Language and Computation* 3.4, pp. 281–332.

Dyvik, Helge (1989). *The PONS Project*. Tech. rep. Department of Linguistics, University of Bergen.

– (2000). "Nødvendige noder i norsk: Grunntrekk i en leksikalsk-funksjonell beskrivelse av norsk syntaks [Necessary nodes in Norwegian: Basic properties of a lexical-functional description of Norwegian syntax]". In: *Menneske, språk og felleskap*. Ed. by Øivin Andersen, Kjersti Fløttum, and Torodd Kinn. Oslo: Novus forlag, pp. 25–45.

Dyvik, Helge, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse, and Martha Thunes (2016). "NorGramBank: A 'Deep' Treebank for Norwegian". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis. ELRA. Portorož, Slovenia, pp. 3555–3562.

Haugereid, Petter (2017). "An incremental approach to gapping and conjunction reduction". In: *Proceedings of the 24th International Conference on Head-Driven Phrase Structure Grammar, University of Kentucky, Lexington*. Ed. by Stefan Müller. Stanford, CA: CSLI Publications, pp. 179–198.

Haugereid, Petter and Mathieu Morey (2012). "A left-branching grammar design for incremental parsing". In: *Proceedings of the 19th International Conference on Head-driven Phrase Structure Grammar, Chungnam National University Daejeon*. Ed. by Stefan Müller. Stanford, CA: CSLI Publications, pp. 181–194.

Oepen, Stephan, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger (2007). "Towards hybrid quality-oriented machine translation. On linguistics and probabilities in MT". In: *In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation*.

Rosén, Victoria, Koenraad De Smedt, Helge Dyvik, and Paul Meurer (2005). "TREPIL: Developing methods and tools for multilevel treebank construction". In: *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*. Ed. by Montserrat Civit, Sandra Kübler, and Ma. Antònia Martí, pp. 161–172.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer, and Helge Dyvik (2012). "An Open Infrastructure for Advanced Treebanking". In: *META-RESEARCH Workshop on Ad-*

vanced Treebanking at LREC2012. Ed. by Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco. Istanbul, Turkey, pp. 22–29.

A word or two?

Christer Johansson

Abstract. The tendency for people to write compounds as two separate words, i.e. decompounding, is, for Scandinavian, often attributed to influence from English. However, English writers also both accidentally compound and decompound words. This article introduces *serendipity* as a statistical signal of surprise, i.e. deviance from expectations. Examples show that this measure can decide many cases of accidental compounding or decompounding by estimating which alternative is over-represented. Interestingly, the least frequent alternative can be clearly over-represented, thus providing a signal that is different from probability estimates, and linked to change in probability.

1 Introduction

People sometimes accidentally miss a key when writing on a *typewriter* (which is one example of an English compound), resulting in 'words' such as *isthere*, and sometimes people hit a space, where it should not be, resulting in words such as *tooth brush* instead of *toothbrush*. Missing a space between two words that are not part of a compound, i.e. *accidental compounding*, can be viewed as a more or less random process that happens at some rate of error. Insertion of an extraneous space is often related to a morpheme boundary, i.e. *accidental decompounding*. This latter process could be a step in approximating where to put a space. Both accidental compounding and decompounding can be a problem for writers, who risk mockery from ungenerous readers.

Accidental compounding: A compound phrase consisting of two separate words are sometimes wrongly written as one word. This could be an accident, such as when two words that often occur together are written as one word: for example *is there* written as *isthere*. Some compound phrases such as *power station* that should be written as two words, are fairly often wrongly written as one word. One explanation is that they are accidental compounds, just like *isthere*.

Accidental decompounding: One explanation for decompounding words such as *lighthouse* is that it is simply a matter of an accidental insertion of a space at a morpheme boundary. There is a background rate of how often we insert a space by mistake at

The very model of a modern linguist edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1329.

such a boundary. However, it is clear that this rate is affected by how often we see the word written correctly and incorrectly. *Football* is almost never decompounded, and *muscle car* is rarely written as one word.

There is less compounding in English than in German or Scandinavian, with many exceptions such as *firefighter*, *football* and *toothbrush*. Sometimes, an English compound begins as a two-word phrase (e.g. *jay walking*, which is now *jaywalking* and the meaning of *jay* in the compound is more or less lost). Such phrases tend to drift towards a one-word compound with increased usage, and correspondingly more specific meaning.

Decompounding words may lead to misunderstandings, as it can affect lexical choice. For example, in Swedish *kassa apparater* are *faulty machines*, but *kassaapparater* are *cashier's registers*. In Swedish it is not common for two vowels to clash, thus an inserted space might reflect the lower transition probability between two vowels inside a word, compared to between words. Language-specific letter transition probabilities have been shown to affect reaction times and accuracy for decision tasks on Norwegian Bokmål and English (Van Kesteren et al. 2012).

The proportion between writing in one or two words, could be more or less surprising. The relative frequency of a one-word compound is typically higher than accounted for by the proportion of words that are accidentally compounded, i.e. the components are not statistically independent.

The new measure of serendipity, as it is introduced in this article, is interesting as an alternative to thinking in the absolute probabilities that most people are not good at estimating. For example, we have a tendency to overestimate the probability of two events occurring together (the Conjunction Fallacy, see section on prerequisites) and also attribute causation to rare events that happen in close sequence (*post hoc ergo propter hoc*, or Causation Fallacy) or repeatedly occur together (correlation implies Causation Fallacy).

What we need is a measure that is insensitive to the number of examples, for example by putting more emphasis on *effect size* rather than *significance* (Johansson 2013). I will also argue that we think more like gamblers, in that we value information that *changes* probabilities more than we value absolute probabilities. If we get information that makes a horse ten times more likely to win, wouldn't we put some money on it even if it still is an unlikely winner? Serendipity is a measure of surprise, and surprise is a good trigger for learning.

In a famous review, Chomsky (1959) draws a caricature of Skinner's research on verbal behavior, by more or less equating the approach with reinforcement learning in animal studies (MacCorquodale 1970). Chomsky's review gave the impression that statistics was not very useful for investigating language structure, and put focus on how improbable a simplistic probabilistic calculation of recursive structures would be.

The question of which signals we use to find linguistic information is not resolved, except by the assumption of innate structures; i.e. we do not find it because it is already there. Optimizing the probability of a sentence is obviously hard given a limited sample, and the infinite possibilities of language to form new sentences and new words. This article will show how the change in probability when comparing alternatives can be used for a seemingly simple task of deciding whether (any) two consecutive words should be written as one or two words. This expands on work by Rømcke and Johansson (2008), where frequencies from a search engine were used to decide categories for named entities, by comparing frequency responses to the words in contexts such as *Hotel in Bergen* or *Her name is Bergen*. Search engine frequencies have also been used to investigate the dative alternation, using frequency responses to the two versions of the dative construction for a set of different dative verbs (Jenset and Johansson 2013). The aim of the examples is to illustrate the signal surprise and expectation, as measured by a new measure.

This article will introduce *serendipity* as the pointwise effect size, by showing how to distribute effect size over the contributions to significance of the individual cells. Crucially, effect size and serendipity are insensitive to how much data we use. This article will begin with some prerequisites, related to statistical independence, cross table testing, significance, effect size and what could be called *serendipity* (i.e., the effect size of a single cell in a cross table), and using Google as a quick and dirty source of observed frequencies.

2 Prerequisites

2.1 Cross tables

A cross table analysis is an analysis of frequencies that tests whether rows and columns are statistically independent of each other. The most basic case is the 2 rows by 2 columns, and it is certainly the easiest cross table to interpret. The null hypothesis is that the rows and columns are independent of each other. Let us consider a simple cross table and calculate the expected independent frequencies of each cell. Let $a + c = R_1$ and $b + d = R_2$ be the total frequencies for row 1 and 2, and $a + b = C_1$ and $c + d = C_2$ the total frequencies of column 1 and 2, and $a + b + c + d = T$ is the total (cf. Table 1).

a	c	R_1
b	d	R_2
C_1	C_2	T

Table 1: A cross table

If the rows and columns are independent then the probability of belonging to row 1 is R_1/T, and R_2/T is the probability of belonging to row 2. The probability of belonging to column 1 is C_1/T and C_2/T is the probability of belonging to column 2.

Assuming independence, the probability of belonging to a cell that is the combination of a row and a column, is simply the multiplication of the row and column probabilities. We can calculate the expected frequencies in each cell (cf. Table 2) by distributing the total by the proportion in each cell.

In order to test for significant deviation from independence, we should look at the difference between observed and expected frequencies (e.g., $O_{11} - E_{11}$); if this difference is positive, that cell is *over-represented* and if it is negative, it is *under-represented*. The test sums up the square of these differences, each one compared with its expected frequency: $\chi^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$; this positive sum indicates how rare it would be to find so much deviance if the rows and columns are indeed independent. In order to look it up in the χ^2-distribution, you need to know how many ways this could happen (i.e. the degrees of freedom). The 2-by-2-table has one degree of freedom, since if you know the value in one cell, you can easily calculate the rest from the row and column sums. This is called degrees of freedom (*df*), because the rest of the cells are uniquely determined by the row and column sums if we know the value for $R - 1$ row cells and $C - 1$ column cells, simultaneously, i.e. $df = (R - 1) * (C - 1)$, where R and C are the number of rows and columns, respectively. It should not be a big surprise if there are significant deviations from independence for language data; after all the process that generated the frequencies (for example, writing an essay) is not a random process. Significance only tells us if it is likely or not that the rows and columns are independent. It does not tell how large or how relevant the effect is.

$E_{11} = \frac{R_1 * C_1}{T}$	$E_{12} = \frac{R_1 * C_2}{T}$
$E_{21} = \frac{R_2 * C_1}{T}$	$E_{22} = \frac{R_2 * C_2}{T}$

Table 2: The expected frequencies

2.1.1 Pearson residuals

If we want to say which cells contribute more to *significance*, then we should look at the *signed* Pearson residuals, which measure the signed contribution to significance in each cell: $\frac{(O_{ij}-E_{ij})}{\sqrt{E_{ij}}}$ from the term $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ in the χ^2-formula, before squaring. One reason for squaring is to sum up deviance in absolute values, and use this to decide the fit of the observations to a model of independence.

2.1.2 Association plots

An association plot is a tool to graphically explore and visualize the effects of each cell in a table. The R command *assocplot* can be used, but the function *assoc* from the

Visualizing Categorical Data (vcd) package (Meyer et al. 2016) provides more possibilities, for example visualizing the Pearson residuals range with a color gradient. The association plot provides bars, whose base is proportional to $\sqrt{E_{ij}}$, and the height is proportional to $O_{ij} - E_{ij}$ and thus the width of the base tells about expectations in that cell, and the height of the bar tells about the deviance from expectations.

2.2 Effect size

Effect size for a χ^2 test is calculated as $\Phi = \sqrt{\frac{\chi^2}{N}}$. Cramér's Φ can be generalized to larger tables, using Cramér's $\nu = \sqrt{\frac{\chi^2}{df*N}}$, where df is the smallest number of the number of either $rows - 1$ or $columns - 1$. The effect size is nearly independent of how many observations the table represents, whereas almost any real difference between observed and expected can be detected with significance by sampling large enough samples from the population. Therefore, if we want to compare results, we should include the effect size. Significance is just a receipt that we have observed a deviance from independence that cannot be explained by random chance.

2.3 Serendipity or the effect size per cell

We are interested in each cell's contribution to the effect size. One way is to note each cell's contribution to the χ^2 statistic compared to the overall χ^2, and this tells each cell's proportional contribution to the effect size measure. The effect size in each cell i, distributed by each cell's contribution to significance (i.e. to χ^2) is given in formula (1).

$$\Phi \sum_{i=1}^{n} (\frac{(O_i - E_i)^2}{E_i})/\chi^2 => \frac{\Phi}{\chi^2}(\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} ... \frac{(O_n - E_n)^2}{E_n}) \quad (1)$$

where Φ and χ^2 are numbers and $\frac{(O_i - E_i)^2}{E_i}$ are terms in a series, which can be ordered in a table.

Proof of correctness for $\chi^2 > 0$: Note $1 = \frac{\chi^2}{\chi^2}$ and rewrite according to definition formula (2).

$$\sum_{i=1}^{n} (\frac{(O_i - E_i)^2}{E_i}/\chi^2) = \frac{1}{\chi^2} \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

which are the terms in the series we need to distribute Φ over all cells.

There are still some problems with the desired function. First of all, division is sensitive when χ^2 is close to zero. This can be fixed by adding one to the numerator and the denominator, and we can let the one in the numerator divide up equally on all the cells. Finally we may scale the value by multiplying by 100 and rounding to two decimals. The scaling is just cosmetic, and makes it easier to read the output. If you sum up the absolute values of all cells you will get back the overall Φ, but remember we have scaled the value by multiplying with 100. The effect size is actually generalized to any size of table by using Cramér's ν to correct for increased degrees of freedom (df).

In the programming language R we define the function in (3) that returns a table with the signed effect size for each cell:

(3)
```
serendipity <-
  function (x){
    df <- min(nrow(x),ncol(x))
    if (df>1) df <- df - 1
    model <- chisq.test(x,correct=F)
    phi <- sqrt(model$statistic/(df*sum(x)))
    o <- model$observed
    e <- model$expected
    s <- sign(o-e)
    phi2 <- phi*((1/prod(dim(x))+ (o-e)^2/e)/(1+model$statistic) )
  return ( round(100*s*phi2, 2) )
  }
```

2.4 Conjunction Fallacy

The probability of two events occurring together (in *conjunction*) is always less than or equal to the probability of either one occurring alone (cf. Wikipedia on Conjunction Fallacy). Tversky and Kahneman (1983) investigated some conditions under which we are more likely to estimate the conjunction as more probable than just one of the events. The classic example presents Linda as having several characteristics of a feminist activist, but nothing to suggest that she is a bank teller. In that situation, most subjects would state that Linda is more likely a feminist and a bank teller, than a bank teller. One possible flaw, is that the alternatives are read in contrast to each other and therefore the alternative that she is a bank teller is actively read as: *she is a bank teller but not a feminist*. This was controlled for in several follow up experiments (ibid.). One version had two explicit arguments to choose from, either argument *a)* "Linda is more likely to be a bank teller than she is to be a feminist bank teller, because every feminist bank teller is a bank teller, but some women bank tellers are not feminists, and Linda could be one of them" (ibid., p. 299) or argument *b)* "Linda is more likely to be a feminist bank teller than she is likely to be a bank teller, because she resembles an active feminist more than she resembles a bank teller" (ibid.). A majority of subjects (65%, 58) preferred alternative *b*, which is still better than the 85% that preferred the conjunction, if there were no explicit arguments for the alternatives. Gould (1988) presents a popular text on this and other statistical fallacies.

2.5 Google frequency estimates

Google provides frequency estimates of search phrases. In my experience, it works better for short phrases, and not too many context words, where examples could be retrieved from big tables rather than estimated. Remember that the frequencies are *estimated*, and may not be accurate. For our purposes we are more interested in the

proportions than in the absolute frequencies. Google frequencies give an estimate of the number of *documents* that contain the search terms, which in itself points to conservative estimates. However, the collection of documents may contain many duplicates. Furthermore, the estimates may have uncertainties that vary non-linearly with the size of the frequency estimate, but there are very few other sources that covers so much of the lower frequency words, which is especially important when we look for low frequency compounds. Obviously, there are data sources of much better quality. There are, however, many reasons why we could prefer Google frequencies. First, when looking for words in context it is crucial to have as many examples as possible. As a comparison, Kuperman and Bertram (2013) finds only 27 examples each of *apple sauce* and *applesauce* in a controlled corpus, whereas Google finds 20 million estimated documents for *applesauce* versus 450 thousand for *apple sauce* (some documents may mention both variants). Second, Google gives document frequencies, which may actually lessen the bias of individual writers, who may overuse certain patterns. Third, Google does not (always) normalize words, so misspelled words or compounds can be represented. Fourth, search through the Google search engine makes replication widely available for almost anybody with Internet access. Finally, Google is updated more frequently than most corpora, which is important when we are interested in contemporary usage. However, it would obviously be good to have access to a linguistic search engine, since Google's search engine is not tailored for the needs of linguists and therefore may prioritize other issues such as bandwidth capacity. The algorithms that are used by Google may also change without notice. It is also an idea to build future applications, where part of the computing is done on the Internet as a distributed system.

In my experience, Google frequencies may often display a machine version of the above mentioned Conjunction Fallacy, i.e. a more specific search may very well indicate more documents rather than fewer. For the normal user of the search engine, this is not a problem as long as the highest ranking documents are the most relevant. For serious research this means that the frequencies should be seen as illustrations rather than hard facts. As will be clear from the examples, in practice the proportions are often very clear, which means that only large errors will affect the decisions based on the effect size measure introduced in this article.

2.5.1 Is it wiki or kiwi?

Table 3 works as an illustration: When searching for the words *kiwi* and *wiki* with and without context words, millions of documents were indicated. Note that *kiwi* is the least frequent alternative both with and without context words. However, when we put on the effect-size goggles, it is clear that we could choose *wiki* if there is no other information (positive effect size = 0.03) and we should choose *kiwi* given the context words *banana* and *fruit* (positive effect size = 15.22), because it is much more frequent than expected. More specifically, we could even program a computer to use the

effect-size measure, as one of many measures, to take decisions between alternatives. It should be clear that the measure is not only probability of occurrence, but directed deviance from expectations, where expectations are set up by statistical independence of (structured) alternatives. Whether people similarly use effect size to guide intuitions about probabilities is a research question that has been hinted at previously, when discussing the Conjunction Fallacy. It should also be noted that part of the work is done by selecting alternatives to compare with, which is one way to establish a baseline for expectations.

	word	+fruit +banana
kiwi	71.2 (-0.21)	6.0 (15.22)
wiki	905.0 (0.03)	6.7 (-1.30)

Table 3: Frequency and (effect size) for *kiwi*/*wiki*.

In probabilistic terms, we would always have a larger *probability* of finding a document containing the word *wiki* than one containing *kiwi*, but *kiwi* has a much stronger association with *banana* and *fruit* than the word *wiki*, which is weakly negatively associated with those words (i.e., most documents that contain *wiki* do not mention *fruit* and *banana*). The **Pearson residual** of *kiwi* in context is 15.9, but if the table is divided by 10 then the cell's effect size is still 14.88, but the Pearson residual is just 5.03, illustrating that the effect size is roughly constant, but the contribution to *significance* varies. Effect size is more relevant than significance, *when we are looking for associations*.

In terms of **Bayesian probability**, given that we have a choice between *kiwi* and *wiki*, the *prior probability*, from the column cells and the column totals, of *kiwi* is $712/(9050 + 712) = 0.073$, and of *wiki* $9050/(9050 + 712) = 0.927$; the probability of *kiwi* given *fruit* and *banana* is $60/(60 + 67) = 0.472$; the probability of *wiki* (i.e., not *kiwi*) given *fruit* and *banana* is $67/(60+67) = 0.528$. The *adjusted probability* of *kiwi* is 0.066, which is lower than its prior probability, and the probability of *wiki* is correspondingly 0.934. This shows that the effect size as an association measure is not just Bayesian probability.

In relation to the Conjunction Fallacy, the association between *kiwi* and *fruit* and *banana* is clearly shown by the ratio between odds with and without information. With the contextual information, 6.0 out of 12.7 (i.e., $6.0 + 6.7$) is for *kiwi*, which gives an odds of 0.472, or expressed as a percentage: 47.2% gives *kiwi* in the specific comparison. Without the contextual information, 71.2 out of 976.2 (i.e. $71.2+905.0$) is for *kiwi*, which gives an odds of 0.072. The odds ratio for *kiwi* is $0.472/0.072 = 6.56$, and similarly for *wiki* $0.528/0.927 = 0.570$. Thus, knowing *fruit* and *banana* increases the chances that it is *kiwi* more than 6 times, while it almost halves the odds that it is *wiki*. So even if it is still unlikely that it is *kiwi*, it would be tempting to choose it — if it

was a bet and the payout is set by the prior probability. This looks similar to how the Conjunction Fallacy works in that having the extra information gives an advantage, in the example above knowing about Linda increases the chances that she is both a bank teller and a feminist much more than it increases the chances that she is a bank teller. What looks like a fallacy might in fact be a more or less innate tendency to value *change in probability* much more than the absolute probability. People could also use this as a communicative strategy: mention only information that changes background knowledge.

2.6 Summary

Effect size distributed per cell is a convenient way of investigating associations in a table. It works well with frequencies, and it is intuitive to understand the concepts in terms of over- and under-represented compared to estimates based on statistical independence of rows and columns. Cells that deviate from expectations are marked clearly.

3 Examples

3.1 To compound or not to compound?

The Google frequency (February 6, 2017) of *there is* is 2390 million against 2.720 million for *thereis*. For *is there* there are 460 million documents and for *isthere* 0.500 million documents. The ratio is between $878:1$ and $920:1$, respectively. The rounded ratio of $1000:1$ will do fine as a baseline, which we can call *is there*. This ratio is likely similar in other languages as well, since it is motivated by the same process of missing a keystroke. However, with less material there is a higher risk that the ratio will be more off. Also, increasing use of better spelling correction may influence the frequencies.

Obviously, for most real applications the missing keystroke rate will have to be estimated for the individual for increased precision, and luckily this should not be very hard to do. It could even be a good idea to estimate more precise measures such as the rate of missing a space between any two specified characters.

3.1.1 Is it firefighter or fire fighter?

Table 4 shows the table for deciding between *firefighter* and *fire fighter*, which has a ratio of $65:1$ in favor of being written as one word. Incidentally, the ratio is almost the same for *fireman*, at $68:1$, it just looks like *firefighter* has a 45% higher frequency. Note that *is there* is preferred as two words, even though we have not explicitly said that it should never be written as one word, and therefore it could *adapt* (by lower effect size) to words like *firefighter*. The important part for the decision is merely which way the effects go.

However, can we be sure that it is not two words? Reasoning from statistical hypothesis testing, we would have to *disprove*, or at least make it unlikely, that the word has *not* been split by mistake. This is a bit trickier. One of the most frequent compound

	word	two words
firefighter	51.5 (88.61)	0.795 (-4.67)
is there	1 (-4.65)	1000 (0.27)

Table 4: Frequency in millions and (effect size) for *firefighter*.

words in English is probably *football*. We could compare that with *firefighter*. Table 5 shows that *firefighter* is indeed not as strong a compound as *football*. Compared to *football*, *fire fighter* cannot be completely ruled out as a two word compound that has undergone *accidental compounding*. For a decision, we then have to compare the effect sizes of *firefighter* as one word in Table 4 (88.6) and *fire fighter* as two words in Table 5 (8.8), and 88.6 clearly wins over 8.8. The first conclusion is that *firefighter* is not likely to be explained as *accidental compounding* and the second is that *firefighter* is a weaker compound than *football*. If we are still unsure we could compare it with a reference word such as *banana peel*, see Table 6. The word *firefighter* is one word, but *banana peel* is strongly a two-word compound, when compared to each other.

	word	two words
firefighter	51.5 (-0.18)	0.795 (8.77)
football	1330 (0.17)	0.408 (-0.51)

Table 5: Frequency in millions and (effect size) for *firefighter* vs. *football*.

	word	two words
firefighter	51500 (0.01)	795 (-0.43)
banana peel	4.2 (-1.29)	403 (55.51)

Table 6: Frequency in thousands and (effect size).

3.1.2 Is it Slotts gate or Slottsgate?

One famous example of decompounding in Norwegian is *Øvre Slottsgate* 'Upper Castle Street'. In Oslo, the street sign actually reads *Øvre Slotts gate*. Investigating the web finds that there is indeed a tentative association between decompounding this street name and documents that also contains the word Oslo, see Table 7.

This is also an example of the statistical Conjunction Fallacy for Google frequencies — adding a demand for an extra keyword ought to give fewer documents, but the search engine has detected a strong association between Oslo and this street name, and the estimate is higher, possibly because the search engine has performed a deeper search. This seems to affect the rarer variant more. For comparison, see Table 8 where

the address is made more specific by adding *Øvre* to the street name. From the table we see very small effect sizes in all the cells, but a small preference for associating *Øvre Slotts gate* with Oslo. The decompounded version is associated with Oslo, and we also know that in Oslo there are street signs that show the decompounded version. Since effect sizes are so small for all versions of *Øvre Slotts gate* the decision could be to trust as it was written. In a practical application, the false alarm rate also need to be kept low, and setting an individual threshold for when to suggest an edit makes sense.

	all	oslo
slotts gate	1520 (-3.93)	6420 (4.17)
slottsgate	256000 (0.06)	236000 (-0.07)

Table 7: Frequency and (effect size) for *Slottsgate* ±Oslo.

	all	oslo
øvre slotts gate	3890 (-0.19)	3890 (0.20)
øvre slottsgate	150000 (0.02)	142000 (-0.02)

Table 8: Frequency and (effect size) for *Øvre Slottsgate* ±Oslo.

4 Analysis

The decision for one word over two words is affected by the baselines for the comparisons. From the examples, we have seen that there are two baselines: one for our expectations for accidental compounding and one for accidental decompounding. In order to show how this works for different proportions, two extremes were chosen as a graphical illustration. One baseline proportion is the accidental compound *thereis*, which occurs once for every thousand occurrences of the correct *there is*. The other extreme is a hypothetical word that should be written as two words but often ends up as one word (e.g., *musclecar*) and that proportion is set at 3 incorrect *onewords* to 1 correct *two word*, which is a pessimistic estimate of accidental decompounding. Note that *musclecar* has the same structure as *football*, i.e., a body part and an object. Most non-compounds have detectably more support as non-compounds, but Figure 1 illustrates that two words could be favored, even under conditions where the baseline itself favors one word $3:1$.

	one word	two words
candidate	a	b
baseline	c	d

Table 9: Baseline matrix

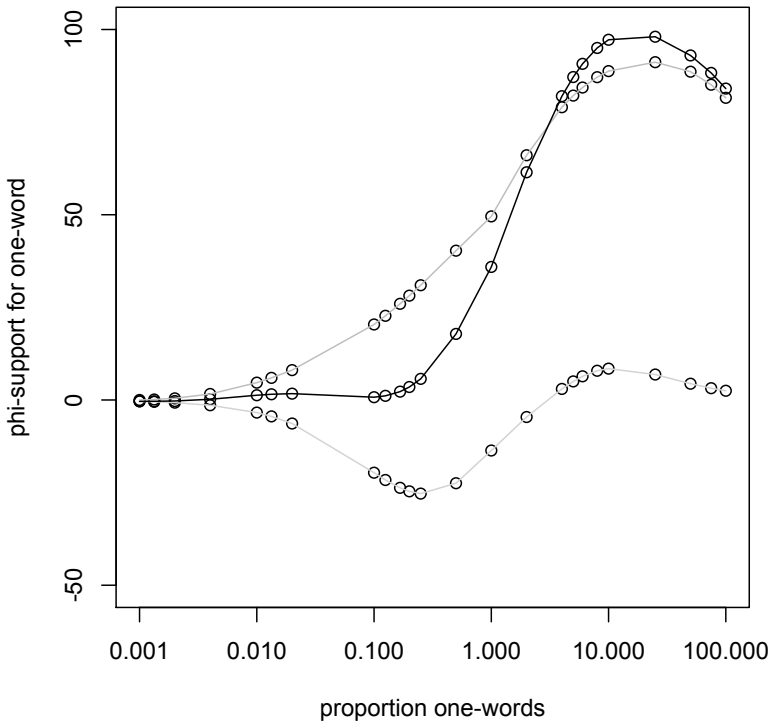


Figure 1: Support for one word. The Y-axis shows the serendipity score and the x-axis shows the proportion $a : b$ (cf. Table 9)

Table 9 shows a matrix for comparing a candidate with a baseline. The frequency for a one-word interpretation is given in the cell marked *a*, and *b* is the two-word frequency. The letters of the table also mark the position for the serendipity scores.

Figure 1 shows how much *phi-support* (i.e., the serendipity score, or pointwise effect size) a new candidate has for being *oneword*. Two different baselines are illustrated: The upper gray line shows a baseline at $c = 1 : d = 1000$ and the lower gray line shows a baseline at $c = 3 : d = 1$.

The score for supporting compounding is calculated for candidate proportions that range from $1 : 1000$ to $100 : 1$. Note that these proportions are plotted on a logarithmic scale on the x-axis.

The lower gray line shows the accumulated serendipity score *against* the one-word interpretations, which accounts for support for two words as well, i.e. the support for a oneword interpretation minus the support for a two word interpretation. The dark line shows the *net support* for one word after accounting for both baselines. The line shows that given these baselines, the positive net support for oneword starts before a $1:1$ proportion, and it quickly gains positive support. After $100:1$ there will be increasingly smaller relative differences between observed and expected frequencies, and less signal for learning, and one way to view this is that there are fewer alternatives and no need to learn as expectations are as observed.

5 Discussion

Are there more linguistically relevant problems, where a measure of association such as the serendipity measure can be used? One possible example is Anaphora Resolution, which is hard to resolve using empirical methods (Nøklestad and Johansson 2005). We found that candidate antecedents can be at long text distances, and most potential candidates are *not* coreferent with the pronoun to resolve. As Nøklestad (2009, p. 215) notes: "Thus, the tendency of the system to classify a candidate as non-antecedent is so strong that a single feature is rarely able to overcome it. This is hardly surprising, given the overwhelming majority of negative examples in the training data (…)".

An idea introduced in this article is that informativeness, and surprise, are related to how much probabilities *change* in a new context, and that this can be used as a trigger for learning. This idea could be applied to coreference resolution: Which antecedent candidate will change the background probability the most? Such an approach has the possibility to find associations that are not the most objectively probable. However, if we take into account that other people may react on, and use, change in probability, this has a good chance to be a relevant signal. Just as the solution to the famous Monty Hall problem (Rosenhouse 2009) lies in realizing that the objective situation that there are two boxes to choose from, has a context and a history that makes it highly rational for a participant to change to the other box, thus changing the initial risk of losing to a chance of winning.

In this article, examples have shown that the risk of both accidental compounding and accidental decompounding has to be taken in account. The serendipity measure that was introduced here reacts on an effect size that is crucially *insensitive* to, or near independent of, the size of the data sample. This means that the measure can be compared, even if we do not know the size of the population. When we compare one-word and two-word 'compounds' with each other, we find that, for English, there seems to be a gliding scale from preferring one word to preferring two words for compounds.

Note that the decision space has not been optimized. For a spelling application, information on *how* something was written should be taken into account. Was the word written fluently, without major hesitations as noticed by time between key presses

(key latencies), or were there several attempts at writing the word? The attempts may have information about the intended word. Are there similar words in the text? It is common to find the intended word correctly spelled in the same text. Keeping track of how often the different kinds of errors occur for a writer could help us discover the optimal point at which more errors are fixed than created. Uncertainty in search engine frequencies is thought to be handled by comparing close examples, with the same number of words in the patterns. The main reason to use search engines is their coverage. Any controlled source with better coverage should be preferred.

In relation to a model of how people handle compounding (Kuperman and Bertram 2013) it is interesting to note that frequency of use, and familiarity, seems to play an important part. As noted previously there is a tendency for unfamiliar or new compounds to start out as spaced compounds (e.g. *jay walker*) and drift towards a fully compounded unit, such as *jaywalker*. Kuperman and Bertram (2013) provide further examples and notice "going against […] orthographic preferences in production comes with a high cost in recognition", which creates a pressure towards adapting to the expectations of readers. They (ibid.) also mention that the strategy for selecting the best alternative form of compounding evolves, as various processes such as morphemic segmentation, semantic integration and visual recognition are influenced by frequency of usage and familiarity. Additionally, there are effects that could be characterized as related to balance between the constituents of the compound; in length, and syllable structure. Such effects may counteract, or support, a transition to more compounding in usage.

Acknowledgement

The author would like to thank two anonymous reviewers for comments that increased the readability, and made some arguments clearer.

References

Chomsky, Noam (1959). "Review of Skinner's *Verbal Behavior*". In: *Language* 35, pp. 26–58.

Gould, Stephen Jay (1988). "The Streak of Streaks". In: *The New York Review of Books*.

Jenset, Gard Buen and Christer Johansson (2013). "Lexical fillers influence the dative alternation: Estimating constructional saliency using web document frequencies". In: *Journal of Quantitative Linguistics* 20.1, pp. 13–44.

Johansson, Christer (2013). "Hunting for significance". In: *The many facets of corpus linguistics in Bergen – In honour of Knut Hofland*. Ed. by Lidun Hareide, Michael Oakes, and Christer Johansson. Vol. 3. Bergen Language and Linguistics Studies (BeLLS) 1. University of Bergen, pp. 211–220.

Kuperman, Victor and Raymond Bertram (2013). "Moving spaces: Spelling alternation in English noun-noun compounds". In: *Language and Cognitive processes* 28.7, pp. 939–966.

MacCorquodale, Kenneth (1970). "On Chomsky's review of Skinner's *Verbal Behavior*". In: *Journal of the Experimental Analysis of Behavior* 13.1, pp. 83–99.

Meyer, David, Achim Zeileis, Kurt Hornik, Florian Gerber, and Michael Friendly (2016). *Package 'vcd' (Visualizing Categorial Data)*. Tech. rep. CRAN.

Nøklestad, Anders (2009). "A machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection". PhD thesis. University of Oslo.

Nøklestad, Anders and Christer Johansson (2005). "Detecting Reference Chains in Norwegian". In: *Proceedings of the 15th Nodalida Conference*. Juensuu, Finland: University of Joensuu electronic publications in linguistics and language technology.

Rømcke, Audun and Christer Johansson (2008). "Named Entity Recognition using the Web". In: *Proceedings of the Second Workshop on Anaphora Resolution*. Ed. by Christer Johansson. Northern European Association for Language Technology (NEALT). Bergen, Norway.

Rosenhouse, Jason (2009). *The Monty Hall problem*. Oxford: Oxford University Press.

Tversky, Amos and Daniel Kahneman (1983). "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement". In: *Psychological Review* 90.4, pp. 293–315.

Van Kesteren, Ron, Ton Dijkstra, and Koenraad De Smedt (2012). "Markedness effects in Norwegian–English bilinguals: Task-dependent use of language-specific letters and bigrams". In: *The Quarterly Journal of Experimental Psychology* 65.11, pp. 2129–2154.

Preserving grammatical functions in LFG

Ronald M. Kaplan

Abstract. Patejuk and Przepiórkowski (2016) have provided arguments and evidence to call into question the traditional role that named grammatical functions have played in the descriptions and representations of Lexical Functional Grammar. They propose reducing the number of distinguished function names to a much more limited set. In this brief paper I examine a few of their observations and find them not yet convincing enough to justify such a fundamental revision of LFG theory. I am also concerned that a less refined structure at the interface between syntax and semantics will only shift to the semantic interpretation component the descriptive and explanatory burden of interpreting idiosyncratic morphosyntactic properties. I conclude that most if not all grammatical function distinctions should be preserved in LFG functional structures.

1 Introduction

Lexical Functional Grammar posits a level of functional structure to decompose the complex mapping between surface word and phrase configurations and the semantic predicate-argument relationships that they express (Kaplan 1989; Kaplan and Bresnan 1982). The f-structure is intended as an intermediate, formal characterization of the syntactic information needed to guide the construction of meaning representations while abstracting away from grammatical details that are semantically irrelevant. The premise of this modular architecture is that the overall form-to-meaning mapping is a nearly decomposable system (Simon 1996) whose apparent complexity can be diminished by a division of labor that separates the correspondence of surface configurations to f-structure from the correspondence of f-structures to semantic representations.

One of the hallmarks of Lexical Functional Grammar from its inception has been the fundamental role that the *names* of grammatical functions play in syntactic descriptions and syntactic representations. The f-structure is defined as a hierarchical attribute-value matrix where symbols like SUBJ, OBJ, and ADJ serve as the attributes that formally identify and distinguish the individual functions. For more than 30 years this fundamental architectural assumption and its associated mathematics have supported precise characterizations of complex grammatical phenomena in a wide variety of languages (see Dalrymple 2001; Bresnan, Asudeh, et al. 2016), the construction of

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1342.

detailed, broad coverage grammars for a more limited set of languages (e.g. Butt et al. 2002), and efficient computational systems for parsing and generation (e.g. Kaplan and Maxwell 1996; Crouch et al. 2008; Wedekind and Kaplan 2012).

However, in a recent provocative paper Patejuk and Przepiórkowski (2016) have called into question that underlying assumption of the LFG architecture. Patejuk and Przepiórkowski (henceforth P&P) argue that most function-name distinctions can and should be eliminated. This is because they are either redundant with other morphosyntactic and semantic properties or because they undercut the analysis of certain well-attested constructions. P&P arrive at a proposal for a bleached-out functional representation with a reduced set of function names consisting only of SUBJ and OBJ and a catch-all DEPS that groups all other clausal entities in an undifferentiated list of HPSG-style 'dependents'.

This is an interesting proposal that certainly deserves more exploration and discussion. In this brief paper I examine some of the syntactic arguments and evidence that Patejuk and Przepiórkowski put forward but find them not yet convincing enough to justify such a fundamental revision of LFG theory. I am also concerned that a less refined structure at the interface between syntax and semantics will only shift to the semantic component the descriptive and explanatory burden of interpreting idiosyncratic morphosyntactic properties. I conclude that most if not all grammatical function distinctions should be preserved in f-structure.

2 The oblique functions

P&P acknowledge that the governable functions SUBJ and OBJ are not directly aligned with particular morphosyntactic properties and therefore have independent theoretical motivation. Setting aside SUBJ and OBJ (and also the ungoverned functions ADJ and XADJ), they point to a deterministic correspondence that is often assumed between syntactic categories and governable grammatical functions for English, as illustrated in (1).

(1)

XP:	NP	PP	CP	InfP (=VP)
GF:	OBJ_θ	OBL_θ	COMP	XCOMP

This picture is more complicated because the OBJ_θ and OBL_θ labels stand for families of functions that are further distinguished in some approaches by values of θ that identify specific thematic roles (e.g. BENEFICIARY or GOAL). These may be flagged in phrase structure by particular prepositions (e.g. *for* or *to*), as in English, or by case markings in languages with richer morphology. P&P argue that the mapping of particular nominals to the proper thematic roles can be achieved without making the θ distinctions in function names.

On any account there must be a specification that correlates particular cases/prepositions with their associated thematic roles (*to* $\longleftrightarrow$ GOAL), and the representation (f-

structure) that serves as the interface between syntax and semantics must encode enough information from the surface configuration so that that specification can be properly interpreted. On the traditional account, that information is extracted from local c-structure properties and converted to explicit, distinctive handles that subsequently give easy access to relevant functional units. This can be accomplished by means of standard functional-description designators in conventional LFG rules and lexical entries as are partially shown in (2).[1]

(2)
$$\begin{array}{lllll} \text{VP} \longrightarrow & \text{V} & \text{NP} & & \text{PP*} \\ & & (\uparrow \textsc{obj})=\downarrow & (\uparrow (\downarrow \textsc{gf})) = (\downarrow \textsc{obj}) & \\ \text{PP} \longrightarrow & \text{P} & \text{NP} & & \\ & & (\uparrow \textsc{obj})=\downarrow & & \\ \textit{to:} & \text{P} & (\uparrow \textsc{gf}) = \textsc{obl}_{\textsc{goal}} & & \end{array}$$

These might characterize f-structure (3) for *John gave a book to Susan.*

(3)
$$\begin{bmatrix} \textsc{pred} & \text{`}\textsc{give}\langle \textsc{subj}, \textsc{obj}, \textsc{obl}_{\textsc{goal}} \rangle\text{'} \\ \textsc{subj} & \begin{bmatrix} \textsc{pred} & \text{`}\textsc{John}\text{'} \end{bmatrix} \\ \textsc{obj} & \begin{bmatrix} \textsc{pred} & \text{`}\textsc{book}\text{'} \end{bmatrix} \\ \textsc{obl}_{\textsc{goal}} & \begin{bmatrix} \textsc{pred} & \text{`}\textsc{Susan}\text{'} \end{bmatrix} \end{bmatrix}$$

Note that under this analysis the f-structure is not cluttered with a separate GOAL/TO feature. The VP rule uses that value to make a local decision about the specific OBL variant, and the result is then recorded as the distinguished grammatical function.

On the account that P&P suggest, the relevant properties of the local configuration presumably would be imported as features into f-structure, perhaps with no motivation other than to enable subsequent discrimination of the units that are collected into an otherwise undifferentiated DEPS list, roughly as in (4).

(4)
$$\begin{bmatrix} \textsc{pred} & \text{`}\textsc{give}\text{'} \\ \textsc{subj} & \boxed{1}\ [\textsc{pred}\ \text{`}\textsc{John}\text{'}] \\ \textsc{obj} & \boxed{2}\ [\textsc{pred}\ \text{`}\textsc{book}\text{'}] \\ \textsc{deps} & \left\langle \boxed{1}, \boxed{2}, \begin{bmatrix} \textsc{pred} & \text{`}\textsc{Susan}\text{'} \\ \textsc{case} & \textsc{goal} \end{bmatrix} \right\rangle \end{bmatrix}$$

1 The Kleene-star asterisk on the PP allows for predicates that subcategorize for multiple co-occurring obliques: *John talked to Susan about the plan.* I also follow the LFG convention that head-marking equations ↑ =↓ are implicit for otherwise unannotated categories.

Here we see that the reduction in the set of function names is accompanied by a compensating increase in f-structure complexity. The CASE feature is explicit in the f-structure and, as P&P propose, the DEPS list redundantly includes the SUBJ and OBJ structures. Apart from the apparent structural complexity, an otherwise unnecessary collection of identification and feature-filtering constraints, essentially another analysis of the space of structures, would also be required to provide a semantic interpretation. Thus, P&P are technically correct in that the correspondence of surface markers and oblique thematic roles can be defined without recourse to these distinguished function names. But the grammatical system may be simpler overall if these distinctions are preserved.

We also see that the semantic form has been reduced to just the predicate name, without the traditional mapping of grammatical functions to semantic arguments. Semantic forms were introduced by Kaplan and Bresnan (1982) as a formal device to encapsulate the syntactic properties of relevance to semantic interpretation while allowing syntactic description to remain agnostic to the details of semantic representation. As P&P and others have noted and as Kaplan and Bresnan anticipated, the syntactic/semantic dependencies that semantic forms encode have been spelled out more explicitly in particular semantic formalisms, e.g. the early Halvorsen and Kaplan (1988) projection architecture and more recently in Glue semantics (Dalrymple 2001). Semantic forms are thus sometimes regarded as redundant with respect to a full-fledged semantic theory. But they are intended to be viewed as succinct characterizations of more elaborate specifications and are designed to support the modularity of the overall grammatical system. Along the same lines, the correspondence between grammatical functions and thematic relations is the province of another relatively independent module within the LFG framework, Lexical Mapping Theory (Levin 1986; Dalrymple 2001; Bresnan, Asudeh, et al. 2016).

3 The function XCOMP

P&P question the independent status of the open complement function XCOMP given its one-to-one correspondence to the category InfP (henceforth VP) that they display in the table in (1). In constructing this table, they have discounted the possibility of assigning XCOMP to adjectival, prepositional, and nominal complements. This is because alternatives to that analysis have appeared in theoretical discussions (Dalrymple, Dyvik, et al. 2004) and in some of the large-scale grammars developed by the Pargram consortium (Butt et al. 2002). Those alternatives (including the PREDLINK proposal) focus mostly on the AP, PP, and NP′ complements of copular constructions, but even then it is recognized that XCOMP is appropriate for at least some non-infinitive examples in some languages (see Dalrymple, Dyvik, et al. 2004 for discussion). Perhaps with less controversy, post-verbal complements as in (5b-c) also show that open complements can be realized by categories other than the VP in (5a).

(5) a. We consider John to be intelligent.
b. We consider John intelligent.
c. We consider John an intelligent manager.

Examples like these are admitted by the rule (6a) and the lexical entry (6b).[2] [3]

(6) a. VP ⟶ V NP VP|AP|PP|NP′
(↑ OBJ)=↓ (↑ XCOMP)=↓

b. *consider:* (↑ PRED) = ‘CONSIDER⟨SUBJ, OBJ, XCOMP⟩’
(↑ XCOMP SUBJ) = (↑ OBJ)

Thus the open XCOMP function cannot be identified only with infinitival phrases in constituent structure, contrary to this particular claim for redundancy that P&P put forward.

4 The function COMP

Turning to the closed complement function, P&P note that COMP is always related to clausal constituents of category CP but the converse is not true: there are CP’s that do not map to COMP. P&P use the paradigm of unlike-category coordination in (7a-c), based on Sag et al. (1985), to make the point. They argue that SUBJ and not COMP should be assigned to CP’s when they stand alone in English pre-verbal positions, given that they can coordinate with uncontroversial nominal subjects.[4] This argument is strengthened by the fact that CP’s can also participate in raising constructions (7d), since then there is no appeal to an indirect inference from coordination.

(7) a. The implications frightened many observers.
b. That Himmler appointed Heydrich frightened many observers.
c. That Himmler appointed Heydrich and the implications thereof frightened many observers.

2 The category NP′ can be derived by a simple type-shifting rule that coerces an ordinary NP into a monadic predicate:

NP′ ⟶ NP
(↑ PRED) = ‘↓⟨SUBJ⟩’

The relation position of the constructed semantic form is filled by ↓, indicating that the semantic interpretation of the entire complement NP is to be taken as a predicate that applies to the controlled subject, just as for complement constructions with other categories.

3 This rule overgenerates in that the verb *consider* does not admit of a prepositional complement:

*We consider John in the park.

See Kaplan and Maxwell (1996), Crouch et al. (2008), and particularly Dalrymple (2017) for discussions of devices that allow individual predicates to restrict the categories that a general phrase structure rule would otherwise allow for their governed functions.

4 Berman (2007) makes a similar argument for German.

d. That Himmler appointed Heydrich seemed to frighten many observers.

P&P appeal to a similar coordination argument to show that CP's can also be mapped to OBJ. That argument is reinforced by several other observations that other researchers have discussed (e.g. Dalrymple and Lødrup 2000; Alsina et al. 2005; Forst 2006). Some post-verbal CP's can undergo passivization, for example, as we see in (8).

(8) a. I believe that the earth is round.
b. That the earth is round was not believed.

But P&P and others also examine evidence for CP's that cannot be assimilated to the SUBJ or OBJ functions. The post-verbal CP in (9a) does not satisfy the conventional passivization test of typical OBJ's.

(9) a. John hoped that it would rain.
b. *That it would rain was hoped.

Dalrymple and Lødrup (2000) suggest preserving the function COMP to label these instances of CP, while P&P follow Alsina et al. (2005) and propose marking these clauses as OBLiques. Forst (2006) also argues for an OBLique account on the basis of considerations from computation and parallel grammar development. Support for this analysis comes from the fact that (non SUBJ or OBJ) finite clauses stand in complementary distribution to traditionally oblique nominals that are marked with particular prepositions/cases.

(10) a. The secretary has already insisted on it. (Forst 2006)
The secretary has already insisted that I have to fill out the form.
b. We weren't aware of the problem. (Alsina et al. 2005)
We weren't aware that Chris yawned.

Forst cites as an advantage of this account that disjunctive subcategorization frames (11a) would no longer be needed for lexical predicates. Only the simpler OBL_{ON} specification for *insist* in (11b) would be required.[5]

(11) a. *insist:* $(\uparrow \text{PRED}) = \text{`INSIST}\langle \text{SUBJ,OBL}_{\text{ON}} \rangle\text{'} \vee (\uparrow \text{PRED}) = \text{`INSIST}\langle \text{SUBJ,COMP} \rangle\text{'}$
b. *insist:* $(\uparrow \text{PRED}) = \text{`INSIST}\langle \text{SUBJ, OBL}_{\text{ON}} \rangle\text{'}$

5 A more compact and possibly more efficient subcategorization frame than (11a) might be expressed with functional uncertainty:

insist: $(\uparrow \text{PRED}) = \text{`INSIST}\langle \text{SUBJ, } \{\text{OBL}_{\text{ON}} \mid \text{COMP}\} \rangle\text{'}$

This pattern can be propagated systematically across the lexicon, perhaps with a general template, as another way of highlighting the complementarity of COMP and obliques.

However, eliminating COMP in favor of predicate-selected obliques (cf. Section 2) may be accompanied by added complexity of the c-structure grammar. The phrase structure rules must be adjusted to anticipate the particular oblique function that a given predicate selects for the CP. This might be done, for example, by a functional uncertainty in the VP rule (12a). Or the annotation on the PP in (2) can be left alone if an unusual exocentric expansion of PP to CP is introduced to guess the particular oblique function in a different way (12b).

(12) a. VP $\longrightarrow$ V { PP* $(\uparrow(\downarrow \text{GF}))=(\downarrow \text{OBJ})$ | CP $(\uparrow\{\text{OBL}_{\text{ON}}|\text{OBL}_{\text{OF}}|\ldots\})=\downarrow$ }

b. PP $\longrightarrow$ {P NP $(\uparrow \text{OBJ})=\downarrow$ | CP $(\uparrow \text{OBJ})=\downarrow$ $(\uparrow \text{GF})\in\{\text{OBL}_{\text{ON}}, \text{OBL}_{\text{OF}}, \ldots\}$ }

There may be other accounts of distributions as in (10a), but their value also must be measured against the impact on other parts of the grammar. As has been suggested, reducing the set of distinguished function-names is not an end in and of itself.

5 The open/closed distinction

P&P argue, as I have indicated, that some grammatical function distinctions are technically unnecessary for syntactic description. They also make a stronger argument, that the distinction between the open complement XCOMP and other closed functions is actually harmful. Their argument is for the most part based on examples of unlike category coordination that also involve open/closed differences in function assignment.

Recall the major premises of the traditional LFG treatment of constituent coordination (Bresnan, Kaplan, et al. 1985; Kaplan and Maxwell 1988; Dalrymple and Kaplan 2000; Dalrymple 2001, and many others): c-structures are derived by substituting a particular category for X in the general metarule (13), the membership annotations map the coordination to a set in f-structure whose elements are the f-structures corresponding to the conjoined constituents, and a so-called 'distributive property' is satisfied by a set if and only if it is satisfied by each of its elements (14).[6]

6 Distribution has typically been defined, in theory and in practice, by simply declaring that some attributes (grammatical functions and morphosyntactic features like CASE) are distributive and others (e.g. PERSON, GENDER, and NUMBER) are not (Kaplan and Maxwell 1996; Crouch et al. 2008; Dalrymple and Kaplan 2000; Dalrymple, King, et al. 2009). Distributive properties are then just those with designators that include distributive attributes.

The Dalrymple and Kaplan (2000) notion of property foreshadows a more general formulation, and Przepiórkowski and Patejuk (2012) propose allowing a larger combination of constraints to be specfied as a unitary distributive property. This would permit in particular arbitrary disjunctive constraints to have narrow scope with respect to coordination, something that has otherwise been encoded indirectly, for example, by using feature decomposition (Dalrymple, King, et al. 2009) or off-path constraints (Przepiórkowski and Patejuk 2012). This idea can be formalized as an explicit operator declaring that an

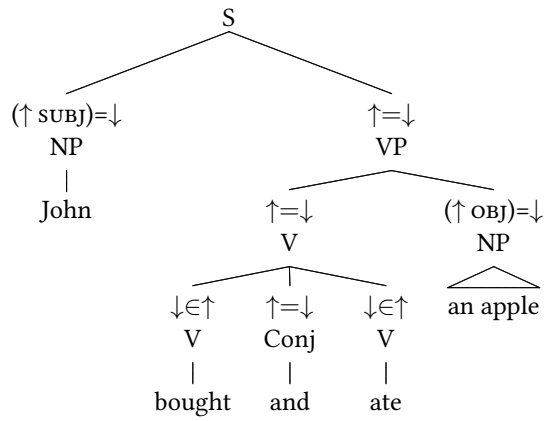
(13) $\begin{array}{lcccc} X & \longrightarrow & X & Conj & X \\ & & \downarrow \in \uparrow & \uparrow = \downarrow & \downarrow \in \uparrow \end{array}$

(14) A structure f satisfies a distributive property P if and only if
f is an f-structure and f satisfies P, or
f is a set and g satisfies P for all g in f.

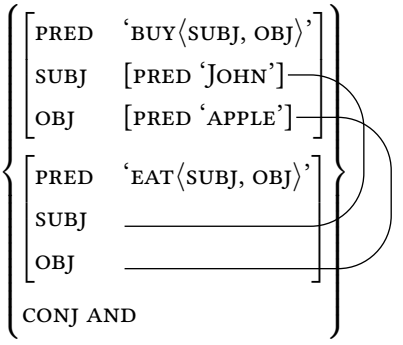
Substituting V for X in (13) will derive the c-structure in (16) for the verb coordination in (15), and it will receive the f-structure (17). Because the set corresponding to the coordinated verb is the head of the VP and S and the grammatical function assignments (SUBJ and OBJ) are distributive, they apply to the f-structures corresponding to each of the verbs. The resulting structure satisfies the subcategorization requirements of each predicate.

(15) John bought and ate an apple.

(16)



(17)



arbitrary description P is a distributive property when it is applied to an f-structure f that happens to be a set:

DISTRIB(f, v, P)

In any invocation (perhaps notated as a built-in template call) f will be a designator (e.g. ↑) and P will be a formula with a variable v that is bound in the scope of P to either the non-set designated by f or to each of its elements in turn.

Of crucial significance, the curved lines in this structure indicate that the two predicates share exactly the same SUBJ and OBJ structures, including the same semantic-form instantiations. Instantiated semantic forms were introduced by Kaplan and Bresnan (1982) to mark for semantic interpretation the difference between two f-structure entities that happen to be described in the same way, and a single entity that serves more than one syntactic function. Thus (15) and (17) contrast with the sentence-level coordination in (18):

(18) a. John bought an apple and John ate an apple.

b.
$$\left\{ \begin{bmatrix} \text{PRED} & \text{`BUY}\langle \text{SUBJ, OBJ}\rangle\text{'} \\ \text{SUBJ} & [\text{PRED `JOHN'}] \\ \text{OBJ} & [\text{PRED `APPLE'}] \end{bmatrix} \begin{bmatrix} \text{PRED} & \text{`EAT}\langle \text{SUBJ, OBJ}\rangle\text{'} \\ \text{SUBJ} & [\text{PRED `JOHN'}] \\ \text{OBJ} & [\text{PRED `APPLE'}] \end{bmatrix} \text{CONJ AND} \right\}$$

Unlike the verb-level coordination in (15), (18) admits the possibility that one apple was bought and another was eaten. Instantiation is the formal device that controls what might otherwise be many other semantic anomalies.[7]

I now return to the question of whether the distinction between open and closed grammatical functions is harmful to syntactic analysis and should therefore be eliminated. P&P base their argument on well-formed examples of unlike-category coordination where distribution would assign an open grammatical function to one of the coordinated phrases and a closed OBJ to the other.

(19) The majority want peace and to live a comfortable life.

The coordination of unlike categories is not in itself a particular problem. The typical approach is to relax the substitution possibilities in the meta-rule (13) so that one of the conjuncts can be realized as a category different from the mother's.

(20) $$\begin{array}{lcccccc} \text{X} & \longrightarrow & \text{X} & & \text{Conj} & & \text{Y} \\ & & \downarrow\in\uparrow & & \uparrow=\downarrow & & \downarrow\in\uparrow \end{array}$$

The match between the mother category and one of its daughters (typically the first as shown here (Peterson 2004), but that issue has not been studied in detail and there

7 In terms of the notions of Glue semantics (see Dalrymple (2001)), the structure (17) provides a single OBJ resource for semantic interpretation with respect to both predicates. Structure (18b) provides two separate resources with accidentally similar properties.

may be wide variation) ensures that at least that daughter satisfies the external categorial requirements of the mother.[8] This relaxation allows for the coordination of unlike closed functions as we saw above (7c) and unlike open functions (21).

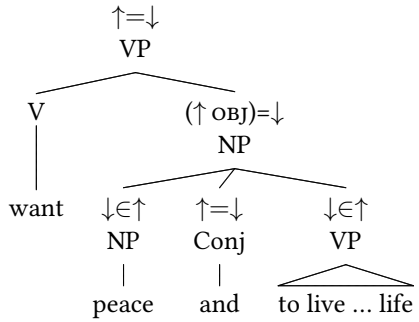
(21) We consider John intelligent and a good manager.

Here the functional control in the lexical entry for *consider* (6b) is distributed to provide a SUBJ for both open elements of the coordination. Combined with the type-shifting rule of footnote 2, this results in an f-structure (22) with appropriate syntactic and semantic properties.

(22)
$$\begin{bmatrix} \text{PRED} & \text{`CONSIDER}\langle \text{SUBJ, OBJ, XCOMP}\rangle\text{'} \\ \text{SUBJ} & [\text{PRED `PRO'}] \\ \text{OBJ} & [\text{PRED `JOHN'}] \\ \text{XCOMP} & \left\{ \begin{bmatrix} \text{PRED} & \text{`INTELLIGENT}\langle \text{SUBJ}\rangle\text{'} \\ \text{SUBJ} & \square \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{`}\begin{bmatrix} \text{PRED} & \text{`MANAGER'} \\ \text{MOD} & \left\{ [\text{PRED `GOOD'}] \right\} \end{bmatrix}\langle \text{SUBJ}\rangle\text{'} \\ \text{SUBJ} & \square \end{bmatrix}, \text{CONJ AND} \right\} \end{bmatrix}$$

Another instance of rule (20) will expand an NP in object position to derive the NP-VP coordination of sentence (19) and assign OBJ to the resulting set.

(23)



This will be compatible with the first of the alternative subcategorization frames for *want* in (24). The problem is that that frame does not provide a SUBJ for *live* in the open VP complement, and the overall f-structure will be incomplete. The other frame

8 Any further predicate-specific restrictions can be imposed by other formal devices, as per the references in footnote 3.

does provide the SUBJ for an XCOMP but does not allow for the OBJ that comes from the c-structure annotation.

(24) *want:* (↑ PRED)='WANT⟨SUBJ, OBJ⟩' ∨ (↑ PRED)='WANT⟨SUBJ, XCOMP⟩'
(↑ XCOMP SUBJ)=(↑ SUBJ)

P&P mention in a footnote that an anonymous reviewer proposed an analysis for sentences like (19) that treats them as instances of non-constituent coordination. I explore that possibility here. Non-constituent coordination has received far less attention in LFG theory than constituent coordination, but a basic framework was laid out by Maxwell and Manning (1996). They introduce systematically a family of new categories and rules just for coordination that subdivide a regular right-side expansion of an ordinary c-structure rule. Those new categories expand so that their concatenation covers the same immediate daughter sequences as the original rule. Consider the VP rule (25) that optionally allows for an OBJ NP and an XCOMP VP.

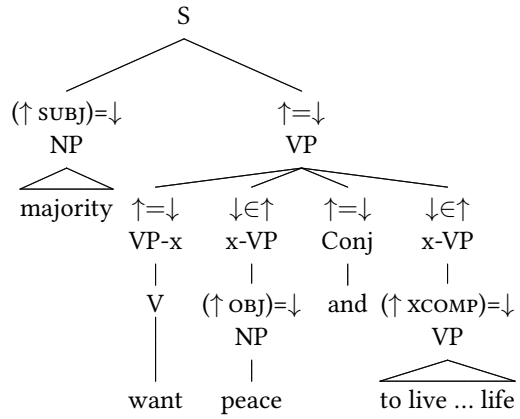
(25) VP ⟶ V (NP) (VP)
(↑ OBJ)=↓ (↑ XCOMP)=↓

According to their proposal, for this case we let x denote the juncture between the initial V and the subsequent optional categories and introduce new categories VP-x and x-VP with expansions as in (26a-b). The alternative VP rule (26c) uses the new categories to cover coordinated VP daughter sequences.

(26) a. VP-x ⟶ V
b. x-VP ⟶ (NP) (VP)
(↑ OBJ)=↓ (↑ XCOMP)=↓
c. VP ⟶ VP-x x-VP Conj x-VP
↑=↓ ↓∈↑ ↑=↓ ↓∈↑

With these rules we can now derive the annotated c-structure (27) for the problematic sentence (19).

(27)



The f-description produced from this c-structure defines a set at the top level (because of the ↑=↓ annotation) that contains two elements. One element has an OBJ corresponding to *peace* and the other has an XCOMP that represents the *live* complement. The disjunctive specification of *want*'s subcategorization requirements (24) still poses a problem. Disjunction in LFG normally has wide scope. Thus either the OBJ frame or the XCOMP frame would be distributed to both elements of the coordination set, and in each case one of the elements will fail the completeness/coherence tests. We must further arrange for the disjunction itself to be distributed and resolved separately on each element. It is well established that functional uncertainties with distributive attributes are independently evaluated on individual set elements, and I make use of that fact to rewrite the *want* lexical entry.[9]

(28) *want:* (↑ PRED) = 'WANT⟨SUBJ, {OBJ | XCOMP (→ SUBJ)=(← SUBJ)}⟩'

This allows *want*'s second argument to be filled by an OBJ in one conjunct and an XCOMP in the other. The subject-control relation is paired as an off-path constraint just with the XCOMP selection: it identifies the XCOMP's SUBJ (designated by (→ SUBJ)) with the matrix SUBJ (designated by (← SUBJ)). With this adjustment we obtain the f-structure (29) for sentence (19).

(29)
$$\left\{\begin{bmatrix} \text{PRED} & \text{'WANT⟨SUBJ, OBJ⟩'} \\ \text{SUBJ} & [\text{PRED 'MAJORITY'}] \\ \text{OBJ} & [\text{PRED 'PEACE'}] \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{'WANT⟨SUBJ, XCOMP⟩'} \\ \text{SUBJ} & \square \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{'LIVE⟨SUBJ, OBJ⟩'} \\ \text{SUBJ} & \square \\ \text{OBJ} & [\text{PRED 'LIFE'}] \end{bmatrix} \end{bmatrix}, \text{CONJ AND} \right\}$$

This non-constituent solution thus assigns appropriate c- and f-structures to (19) while preserving the open/closed complement distinction. As John Maxwell (p.c.)

9 Alternatively, we can declare the disjunctive entry for *want* (24) as a narrow-scope distributive property using the DISTRIB notation proposed in footnote 5:

want: @DISTRIB(↑, V, (V PRED) = 'WANT⟨SUBJ, OBJ⟩'
∨
(V PRED) = 'WANT⟨SUBJ, XCOMP⟩'
(V XCOMP SUBJ) = (V SUBJ))

Indeed, it may be worth exploring whether subcategorization frames and other core lexical constraints should be interpreted distributively as a general convention.

notes, the clearer case of non-constituent coordination in (30) offers further support for this analysis.

(30) The majority want peace on some days and to live a comfortable life on others.

P&P dismiss this approach, however, on semantic grounds. They point to well known observations about the distribution of quantification over coordination (e.g. Partee (1970)), noting the difference in possible interpretations for the single indefinite NP external to a phrasal coordination (31a) compared to a repetition of quantified NPs in a sentence-level coordination (31b).

(31) a. A majority want peace and to live a comfortable life.

b. A majority want peace and a majority want to live a comfortable life.

The same majority is involved in both (31a) events while (31b) admits of two distinct majorities. P&P suggest that a complicated syntax-semantics mapping would be required to distinguish the intended readings of these sentences, given the similarity of their f-structures. But the f-structure for (31a) has the upper SUBJ-to-SUBJ linking line that (29) has for (19). This encodes the fact that a single semantic resource is a participant in both clauses. Crucially, that link is missing in (32), the f-structure for (31b).

(32)
$$\left\{\begin{bmatrix} \text{PRED} & \text{`WANT}\langle\text{SUBJ, OBJ}\rangle\text{'} \\ \text{SUBJ} & [\text{PRED `MAJORITY'}] \\ \text{OBJ} & [\text{PRED `PEACE'}] \end{bmatrix} \right.$$
$$\left.\begin{bmatrix} \text{PRED} & \text{`WANT}\langle\text{SUBJ, XCOMP}\rangle\text{'} \\ \text{SUBJ} & [\text{PRED `MAJORITY'}] \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{`LIVE}\langle\text{SUBJ, OBJ}\rangle\text{'} \\ \text{SUBJ} & \\ \text{OBJ} & [\text{PRED `LIFE'}] \end{bmatrix} \end{bmatrix} \right\}$$
$$\text{CONJ AND}$$

The syntactic representation of shared/unshared resources thus marks a difference that can support the alternative readings. Note also that these semantic differences are orthogonal to the distinctions between open and closed functions, like and unlike category coordination, and constituent and nonconstituent coordination: the sentences (33) exhibit the same semantic contrasts.

(33) A majority want to make money and to live a comfortable life.

A majority want to make money and a majority want to live a comfortable life.

P&P argue from examples like (19) that it is not helpful, and even harmful, to discriminate between open and closed complements. At least for these examples we have seen that this is not the case. Treating this as an instance of nonconstituent coordination, our analysis maintains that functional distinction but still assigns representations that are plausible with respect to both syntax and semantics.

6 Conclusion

I have surveyed some of the arguments and some of the evidence that Patejuk and Przepiórkowski (2016) have presented as motivation for reducing the inventory of grammatical functions that may populate an LFG f-structure. It would be surprising if there were no connection between specific grammatical functions and other morphosyntactic properties, since those properties of words and phrases are what signal those functions in particular configurations. But contrary to P&P and even though it may be technically possible, I have suggested that the overall grammatical system will not be improved if obliques are no longer differentiated or if the open and closed complement functions are collapsed together or with other functions. The denatured representation that P&P propose as a replacement for an articulated f-structure may simplify the syntactic component of the grammatical system at the expense of redundancy and complexity in semantic interpretation. Distinguished grammatical functions abstract away from variation in morphosyntactic detail, preserving (or creating) formal distinctions at the intermediate f-structure level intended to support an overall simpler, modular mapping from surface form to meaning.

Acknowledgments

I thank Agnieszka Patejuk and Adam Przepiórkowski for focusing new attention on some of the fundamental assumptions of LFG theory. Their thoughtful reexamination of basic architectural issues served as the starting point for the present paper. I also benefited from interactions with other participants at the 2016 Headlex meeting that Agnieszka and Adam organized and hosted in Warsaw. I am also indebted to Agnieszka, Adam, Mary Dalrymple, John Maxwell, and the anonymous reviewers for helpful comments on earlier drafts of this paper. Finally, I want to express my gratitude to Helge Dyvik for his enduring friendship and intellectual stimulation over many years.

References

Alsina, Alex, Tara Mohanan, and K. P. Mohanan (2005). "How to Get Rid of the COMP". In: *On-Line Proceedings of the LFG2005 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Berman, Judith (2007). "Functional Identification of Complement Clauses in German and the Specification of COMP". In: *Architectures, Rules, and Preferences: Variations*

on Themes by Joan W. Bresnan. Ed. by Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning. Stanford: CSLI Publications, pp. 69–83.

Bresnan, Joan, Ronald M. Kaplan, and Peter G. Peterson (1985). "Coordination and the Flow of Information Through Phrase Structure". Unpublished manuscript, Xerox PARC.

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler (2016). *Lexical-Functional Syntax*. 2nd ed. Oxford: Wiley Blackwell.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer (2002). "The Parallel Grammar Project". In: *Proceedings of the International Conference on Computational Linguistics (COLING2002) Workshop on Grammar Engineering and Evaluation*. International Committee on Computational Linguistics. Taipei, pp. 1–7.

Crouch, Dick, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula S. Newman (2008). *XLE Documentation*. Palo Alto Research Center. Palo Alto, CA.

Dalrymple, Mary and Ronald M. Kaplan (2000). "Feature Indeterminacy and Feature Resolution". In: *Language* 76.4, pp. 759–798.

Dalrymple, Mary and Helge Lødrup (2000). "The Grammatical Functions of Complement Clauses". In: *On-Line Proceedings of the LFG2000 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Dalrymple, Mary (2001). *Lexical Functional Grammar*. Syntax and Semantics 34. New York: Academic Press.

Dalrymple, Mary, Helge Dyvik, and Tracy Holloway King (2004). "Copular Complements: Closed or Open?" In: *On-Line Proceedings of the LFG2004 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Dalrymple, Mary, Tracy Holloway King, and Louisa Sadler (2009). "Indeterminacy by Underspecification". In: *Journal of Linguistics* 45, pp. 31–68.

Dalrymple, Mary (2017). "Unlike phrase structure category coordination". In: *The very model of a modern linguist: in honor of Helge Dyvik*. Ed. by Victoria Rosén and Koenraad De Smedt. Bergen Language and Linguistics Studies (BeLLS) 8, pp. 33–55.

Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, eds. (1995). *Formal Issues in Lexical-Functional Grammar*. Stanford: CSLI Publications.

Forst, Martin (2006). "COMP in (Parallel) Grammar Writing". In: *On-Line Proceedings of the LFG2006 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Kaplan, Ronald M. and John T. Maxwell III (1996). *LFG Grammar Writer's Workbench*. Palo Alto Research Center. Palo Alto, CA.

Halvorsen, Per-Kristian and Ronald M. Kaplan (1988). "Projections and Semantic Description in Lexical-Functional Grammar". In: *Proceedings of the International Con-*

ference on Fifth Generation Computer Systems (FGCS-88). ICOT. Tokyo, pp. 1116–1122. Reprinted in Dalrymple, Kaplan, Maxwell, and Zaenen (1995, pp. 279–292).

Kaplan, Ronald M. and John T. Maxwell III (1988). "Constituent Coordination in Lexical-Functional Grammar". In: *Proceedings of the International Conference on Computational Linguistics (COLING88)*. International Committee on Computational Linguistics, pp. 303–305. Reprinted in Dalrymple, Kaplan, Maxwell, and Zaenen (1995, pp. 199–210).

Kaplan, Ronald M. (1989). "The Formal Architecture of Lexical-Functional Grammar". In: *Journal of Information Science and Engineering* 5, pp. 305–322. Reprinted in Dalrymple, Kaplan, Maxwell, and Zaenen (1995, pp. 7–22).

Kaplan, Ronald M. and Joan Bresnan (1982). "Lexical-Functional Grammar: A Formal System for Grammatical Representation". In: *The Mental Representation of Grammatical Relations*. Ed. by J. Bresnan. Cambridge, Mass.: MIT Press, pp. 173–281. Reprinted in Dalrymple, Kaplan, Maxwell, and Zaenen (1995, pp. 29–130).

Levin, Lori S. (1986). "Operations on Lexical Forms: Unaccusative Rules in Germanic Languages". PhD thesis. Massachusetts Institute of Technology. Reprinted by Garland Press, New York, 1988.

Maxwell III, John T. and Christopher D. Manning (1996). "A Theory of Non-Constituent Coordination Based on Finite-State Rules". In: *On-Line Proceedings of the LFG96 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.

Partee, Barbara Hall (1970). "Negation, conjunction, and quantifiers: Syntax and semantics". In: *Foundations of Language* 6, pp. 153–165.

Patejuk, Agnieszka and Adam Przepiórkowski (2016). "Reducing Grammatical Functions in LFG". In: *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*. Ed. by Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King, and Stefan Müller. Stanford: CSLI Publications.

Peterson, Peter G. (2004). "Coordination: Consequences of a Lexical-Functional Account". In: *Natural Language and Linguistic Theory* 22.3, pp. 643–679.

Przepiórkowski, Adam and Agnieszka Patejuk (2012). "On case assignment and the coordination of unlikes: The limits of distributive features". In: *On-Line Proceedings of the LFG2012 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications, pp. 479–489.

Sag, Ivan A., Gerald Gazdar, Thomas Wasow, and Steven Weisler (1985). "Coordination and How to Distinguish Categories". In: *Natural Language and Linguistic Theory* 3.2, pp. 117–171.

Simon, Herbert A. (1996). "The Architecture of Complexity: Hierarchic Systems". In: *The Sciences of the Artificial*. 3rd ed. Cambridge, Mass.: M.I.T. Press, pp. 183–216.

Wedekind, Jürgen and Ronald M. Kaplan (2012). "LFG Generation by Grammar Specialization". In: *Computational Linguistics* 38.4, pp. 867–915.

Norwegian *masse*: from measure noun to quantifier

Torodd Kinn

Abstract. For a little more than a century, a new quantifier has been developing in Norwegian: *masse* 'a lot, lots, many, much'. The article compares the quantifier to its source noun *masse* 'matter, mass, large amount'. The historical development is studied based on several corpora. The development of a new quantifier is seen in the larger picture of the variability of measure noun constructions and the tendency for certain kinds of measure nouns to grammaticalize into quantifiers.

1 Introduction

In spoken and informal written Norwegian, a new quantifier has been developing for a few generations, apparently since the decades around 1900. The newcomer *masse* 'a lot, lots, many, much' is advancing into the territory of the older quantifiers *mange* 'many' and *mye* 'much'. Examples (1) and (2) show its use with a count and a noncount noun, respectively, while (3) illustrates that it can also be used as a quantifying adverbial:[1]

(1) *Jeg hadde drukket masse halvlitere*
I had drunk a.lot pints
'I'd drunk lots of pints'

(2) *Alle bruker masse tid på å bevise at Gud fins*
all use a.lot time on to prove that God exists
'Everybody spends lots of time proving that God exists'

(3) *Terry snakker masse om deg*
Terry talks a.lot about you
'Terry talks a lot about you'

1 Sources of examples are provided after the main text. In the interlinear glosses, I use *a.lot* to translate the quantifier *masse* 'a lot, lots, many, much' and *lot* to translate the noun *masse* when it means '(a) lot', alternatively *mass* when appropriate. Morphological abbreviations are kept to a minimum; the following are used when relevant: ABU = plural of abundance; C = common gender; M = masculine; N = neuter; PL = plural; PST = past tense; SG = singular; REFL = reflexive.

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1327.

The origin of the quantifier *masse* is well known and quite transparent; it is the indefinite singular of the masculine noun *masse* 'matter, mass, large amount' used as a measure noun with the meaning 'large amount' (see Section 2). The use of this measure noun is illustrated with examples (4)–(6) parallel to (1)–(3):

(4) *Jeg har truffet en masse mennesker*
I have met a.M lot(M) human.beings
'I've met lots of people'

(5) *Det kan i hvert fall forårsake en masse hat*
it can in every case cause a.M lot(M) hatred
'At any rate, it can cause lots of hatred'

(6) *De hadde spurt en masse og glodd nysgjerrig på ham*
they had asked a.M lot(M) and stared curiously on him
'They'd asked about lots of things and stared curiously at him'

Here, the only apparent difference between the quantifier and the noun is the use of the agreeing indefinite article *en* 'a'. But we will see that there are other differences as well, which firmly establish the status of bare *masse* as a quantifier rather than a noun.

This article investigates the development of the new quantifier from a noun: How and when did it happen, and what is the reason for it? The analysis offered builds crucially on the semantics of the constructions involved, since the observed development needs to be understood as reanalysis that overrides overt morphosyntax.

Below, I will mostly write $masse_Q$ for the quantifier, $masse_{MN}$ for the noun in its measure-noun use/meaning, and $masse_N$ for the noun when it is not a measure noun (see Section 2) or when it is not essential to differentiate between measure noun and non-measure noun.

2 Preliminaries

When an expression like *masse* develops historically from a noun into a quantifier, it crosses a major semantic divide: Whereas nouns designate conceptual things[2] (nominal entities), quantifiers designate conceptual relationships (relational entities). The change involves a significant semantic and syntactic restructuring.

The things designated by nouns are of three fundamental types: individuals (singular count nouns), count masses (plural count nouns), and noncount masses (noncount nouns). Many quantifiers combine with either plural count nouns or noncount nouns and specify the quantity of the count or noncount masses as wholes. Quantifiers meaning 'one', 'every' and some others combine only with singular count nouns. Quantifiers are in many ways similar to adjectives. But the latter combine freely with

2 *Thing* is a term in the conceptual semantics of Cognitive Grammar.

all three types of nouns and specify some quality of individuals (as designated by singular count nouns or as members of the masses designated by plurals) or of arbitrary submasses of noncount masses. For instance, in *three black cats*, the quantifier specifies the cardinality of the count mass and the adjective specifies the colour of the members of that mass. And in *much black coffee*, the quantifier specifies loosely the volume of the noncount mass and the adjective specifies the colour of (the mass and) any given submass.

Measure nouns are a subclass of nouns. They are identified on the basis of their participation in measure noun constructions, also known as pseudopartitives (e.g. Kinn 2001) and under various other terms (cf. Brems 2011, p. 19–26), e.g. (7)–(8).

(7) *en* *mengde* *bøker*
a.M lot(M) books
'a lot of books'

(8) *noen* *glass* *med* *øl*
some.PL glasses with beer
'some glasses of beer'

These are binominal constructions, with a substance noun providing a mostly qualitative categorization of a referent and a measure noun contributing mostly quantitative information about the same referent – plus possibly some case or prepositional marking connecting the nouns (see below). In this article, I will speak about measure nominals and substance nominals as separate parts of measure noun constructions, although one of them will always be part of the other, depending on which noun heads the construction.

Faarlund et al. (1997, p. 238) make a useful distinction between secondary and primary measure nouns: Secondary measure nouns have a relatively clear qualitative meaning in addition to that of quantity, indicating shape (e.g. English *slice, drop*), configuration (*pile, herd*), or containment (*glass, barrel*). Primary measure nouns have more or less exclusively quantitative meaning: specific number (*million, dozen)*, indefinite number (e.g. *number* in *a number of books*), conventional measures (*mile, litre, ton*), indefinite quantity (e.g. *amount* in *a large amount of sugar*). Some measure nouns are restricted to constructions where the substance noun is countable, while others are not. Norwegian $masse_{MN}$ is a primary measure noun of indefinite (large) quantity without any restrictions on the countability of the substance nominal, as shown above by (4) and (5).

Norwegian count nouns regularly exhibit paradigms with four inflectional forms (singular vs. plural and indefinite vs. definite). But measure nouns capable of referring to large quantities are also characterized by the formation of an additional inflectional

form, the abundance plural (Enger and Conzett 2016; Kinn 2004, 2005). Thus, Norwegian Bokmål *masse*$_{MN}$ has the forms *masse* (sg. indef.), *massen* (sg. def.), *masser* (pl. indef.), *massene* (pl. def.), *massevis* (abundance plural).

As illustrated in examples (4) and (5) above, the measure noun and the substance noun in Norwegian measure noun constructions are often juxtaposed, with no marking of one noun being subordinate to the other. This is different from English, where most measure noun constructions involve the use of the preposition of (e.g. *two pounds of sugar, lots of people*).[3] Koptjevskaja-Tamm (2001) shows that European languages commonly exhibit three kinds of marking in measure noun constructions: zero (juxtaposition), prepositional marking of the substance nominal (as in English), and case marking of the substance nominal (e.g. most Slavic languages). In some languages, two or more patterns exist alongside one another, partly in competition. This is the case in Norwegian, where besides juxtaposition there are constructions involving the prepositions *med* 'with' and *av* 'of' (Kinn 2001). There is much variation, depending mostly on properties of the measure nominal: noun meaning, inflectional form, and modification (see further Section 3 for the case of *masse*$_{MN}$).

3 The measure noun *masse*

Derived from a verb meaning 'knead', the Ancient Greek noun *mâza* 'barley-bread, cake' was borrowed into Latin as *massa* 'lump, dough, bulk (of material)'. This word is found in various forms in European languages, including Spanish (*masa*), French (*masse*), English (*mass*), and German (*Masse*), in Swedish and Dutch (*massa*) and in Danish and Norwegian (*masse*). Into Danish, which was the written language of Norway for several centuries, it was borrowed as *Massa*, a form that was gradually replaced by *Masse*, cf. (9) and (10):

(9) *Det er en Gift af en ubekiendt Natur, som omløber i*
it is a poison of an unknown nature which runs.around in
Blodets Massa
the.blood's mass
'It is a poison of an unknown nature which circulates in the blood mass'

(10) *I en saa uhyre Sal med en saadan Masse Mennesker er*
in a so immense hall with a.C such.C mass(C) human.beings is
det ligemeget …
it as.much …
'In such an immense hall with such a lot of people, it does not matter … '

According to the modern dictionary *Bokmålsordboka* (*Bokmålsordboka* 2005, s.v. masse), *masse*$_{N}$ now has four main meaning variants: (1) '(shapeless) matter, sub-

3 Numeral nouns are partly exceptions to this, e.g. *two million people*, but *millions of people*.

stance', (2) 'mass' (the physics notion), (3) 'large amount', and (4) 'most people, the masses' (when used in the definite plural).[4] The first and second variants are illustrated in (11)–(12):

(11) *en skvulpende, seig masse som kalles flytende metallisk hydrogen*
a.M sloshing viscous.M matter(M) that is.called floating metallic hydrogen
'a sloshing, viscous matter called liquid metallic hydrogen'

(12) *Bruk grafen til å bestemme solas masse*
use the.graph to to decide the.Sun's mass
'Use the graph to decide the mass of the Sun'

The third variant mentioned in *Bokmålsordboka* ('large amount') may be classified as a measure noun, and it is from this that $masse_Q$ has developed. The noun is frequent as the head of compounds, e.g. *muskelmasse* 'muscle mass', *kokosmasse* 'shredded coconut', *fugemasse* 'grout' (lit. 'joint mass'), *folkemasse* 'crowd of people'.

As noted above, Norwegian measure noun constructions may be juxtapositional or employ either of the prepositions *med* 'with' and *av* 'of'. Since $masse_Q$ has developed from indefinite singular *en masse*, the use of juxtaposition or a preposition after the indefinite singular $masse_{MN}$ is more central here than that seen with other forms of $masse_{MN}$. Indefinite singular *en masse* (without further modification, or modified by an intensifying adjectival expression, typically *hel* 'whole' or *helvetes* 'helluva') is usually used in juxtaposition, illustrated in (4)–(5). The preposition *med* is sometimes used, as in (13), while the use of *av* is mostly restricted to contexts with other meanings of $masse_N$. But when $masse_{MN}$ is (uncharacteristically) modified by a dimensional adjective, *av* still tends to be used, as in (14); the borderline between measure noun and other uses is here often quite fuzzy.

(13) *Dette kommer selvfølgelig til å koste en masse med penger*
this comes of.course to to cost a.M lot(M) with money.PL
'This is going to cost lots of money, of course'

(14) *[De] oppleves som … problematiske for en stor masse av samtidige lesere*
they are.experienced as … problematic for a.M large.M lot(M) of contemporary readers
'They are felt as problematic for a large number of contemporary readers'

4 Variant (4) might better be regarded as a version of variant (3), but will not be discussed further here.

Turning to the external agreement properties of measure noun constructions with *masse*$_{MN}$, it should be noted that it is the substance noun rather than *masse*$_{MN}$ that determines agreement on adjectival predicate complements and predicate adjuncts. Two examples are provided in (15)–(16), where the adjectives (*verdiløse* and *kunnskapsmette*) agree with the substance nouns (*penger* and *studiner*, respectively). Agreement with *masse*$_{MN}$ (verdiløs, kunnskapsmett) would, in my judgement, be ungrammatical.

(15) *En viktig sideeffekt ... er at en masse svarte penger*
an important side.effect ... is that a.M lot(M) black.PL money.PL
blir verdiløse
become worthless.PL
'An important side effect is that lots of black money becomes worthless'

(16) *En masse yndige studiner veltet kunnskapsmette ... inn*
a.M lot(M) graceful.PL female.students crowded knowledge.full.PL ... in
på bussen
on the.bus
'Lots of graceful female students crowded into the bus, their heads packed with knowledge.'

Note that even if one inserts the preposition *med*, the adjective agrees with the substance noun; it is hard to find authentic examples, though. Using the preposition *av* does not seem natural in these examples.

4 The quantifier *masse*

In order to find early instances of *masse*$_{Q}$, I have searched in the collections of the National Library of Norway.[5] I may have overlooked examples, but the oldest case of *masse*$_{Q}$ that I have found is from a book translated from English, published in 1886. The quantifier is capitalized in agreement with its nominal origin and the orthography of 19th century Danish. The example is given in (17).

(17) *Træstammen ... stod midt i Masse af halvraadne Stubber*
the.tree.trunk ... stood in.the.middle in a.lot of half.rotten stumps
og Rødder
and roots
'The tree trunk stood among lots of half rotten stumps and roots'

It may be noted that the quantifier is followed by the preposition *af* (modern Norwegian *av*) 'of', which sounds slightly strange in (modern) Norwegian but is apparently the normal use of *masse*$_{Q}$ in modern Danish (see below). In the next example that I

5 http://www.nb.no/

have found (also in a book translated from English) from 1907, *masse*$_Q$ is followed by *med* 'with', see (18); this sounds acceptable in modern Norwegian, too.

(18) *Nei tak, maa jeg be om noget lyst og livligt og*
no thank may I pray about something bright and lively and
masse med sol!
a.lot with sun
'No thanks, may I ask for something bright and lively and lots of sun!'

Example (19) is from a book published in 1913, containing students' songs from the period 1813–1913. The book does not tell the age of this particular song, but it refers to a "children's help day", a phenomenon occurring first in Kristiania (now Oslo) in 1906, which narrows the range of possible periods for the expression to 1906–13.

(19) *Saa næste gang de masse smaa skal hjælpes, vil jeg*
so next time the.PL a.lot small.PL shall be.helped will I
passe paa at faa en tiggerbøsse selv at drasse paa
make.sure on to get a beggar.box self to haul on
'So the next time the many small ones are going to be helped, I will make sure to have a beggar's box to haul around myself'

Here, *masse*$_Q$ is used in a definite noun phrase, a usage that appears to have gone extinct; at least, this is the only instance I have found of it, and it sounds strange to the modern speaker.

The oldest example that I have found of the typical use of *masse*$_Q$ — in indefinite noun phrases without a following preposition — is from 1914 and used in a Norwegian novel, see (20). The next two, (21)–(22), are from translations from English and Swedish and published in 1916 and 1919, respectively.

(20) *og Hans kommer hjem med masse skiddent tøi*
and Hans comes home with a.lot dirty laundry
'and Hans comes home with lots of dirty laundry'

(21) *posten kom; med masse kort, pakker og brever*
the.post came with a.lot cards packets and letters
'the post arrived, with lots of cards, packets and letters'

(22) *den lignet mest en liten dverg med masse rynker og stort,*
it resembled most a little dwarf with a.lot wrinkles and large
sort skjeg
black beard
'it resembled most (of all) a little dwarf with lots of wrinkles and a large, black beard'

It would seem that the use of *masse* as a quantifier started to become conventionalized around 1900. Provided that the example from 1886 is not just a misprint, the development towards a quantifier had already started by then, and it is hard to estimate exactly when it began.

Norsk riksmålsordbok (1937-1957, vol. 2, part 1, s.v. masse I), whose first issues were edited before World War II, states that $masse_{MN}$ (rather than the other meanings of $masse_{N}$) belongs to "familial" language. Further, it is noted that it may be used "uten ubest[emt] artikkel, følt som adj[ektiv]" — 'without the definite article, felt to be an adjective' (recall the semantic resemblance between adjectives and quantifiers, modifying different aspects of nominal meaning). One example of such usage is given in (23).

(23) *han har hatt masse penger*
he has had a.lot money.PL
'he must have had lots of money'

In *Norsk referansegrammatikk* (Faarlund et al. 1997, p. 238) it is observed half a century later that *masse* may be used without the indefinite article *en* 'a', achieving "nærmest ren kvantorstatus" — 'almost a pure quantifier status'.

It may be noted that the development of $masse_{MN}$ into a quantifier is not an isolated Norwegian phenomenon, but is also found in Swedish and Danish. Swedish $masse_{Q}$ is like Norwegian $masse_{Q}$ in normally being immediately followed by the substance noun, while Danish $masse_{Q}$ tends to be followed by *af* 'of', cf. (24) and (25), respectively.[6]

(24) *Kände hur massa stearin rann på ryggen när jag sjöng*
felt how a.lot stearin ran on the.back when I sang
'(I) felt how lots of candle wax was running down my back as I was singing'

(25) *Et velholdt feriehus med masse af charme*
a well.kept holiday.house with a.lot of charm
'A well kept holiday house with lots of charm'

See also Clerck and Brems (2015) for the grammaticalization of *mass(es)* of in English.

Being a noun, $masse_{MN}$ is typically preceded by the agreeing indefinite article *en* and sometimes an agreeing adjective. Quantifiers, on the other hand, resemble adjectives semantically and may take degree modifiers if their semantics is suitable for that. Thus, while $masse_{MN}$ may be modified by the agreeing adjective *enorm* 'enormous' in (26), $masse_{Q}$ may be modified by the same adjective in the neuter singular form *enormt* 'enormous(ly)' as in (27); this form is the one that adjectives take when used adverbially.

6 I have not investigated the frequencies of these quantifiers.

(26) *de har en enorm masse nyttig informasjon*
they have a.M enormous.M lot(M) useful information
'They have an enormous amount of useful information'

(27) *Lenken gir også tilgang til enormt masse info*
the.link gives also access to enormous.N a.lot information
'The link also gives access to an enormous amount of information'

In the oldest corpus that I have used (cf. Section 4), the demonstrative adjective *saadan* 'such' (modern: *sånn*) is used in front of $masse_{MN}$, as in (28), showing the nominal status of *masse*. Modern $masse_{Q}$ is preceded by the demonstrative adverb *så* 'so', as in (29), demonstrating the change from measure noun to quantifier:

(28) *jeg skrev en saadan Masse Breve til ham og Broderen om*
I wrote a.M such.M lot(M) letters to him and the.brother about
alverdens Smaating
all.the.world's little.things
'I wrote such a lot of letters to him and his brother about all kinds of little things'

(29) *da så hun så masse rare ting*
then saw she so a.lot strange things
'then she saw so many strange things'

5 A corpus study of *masse* as a measure noun and as a quantifier

In order to look closer into the development of $masse_{Q}$ through time, I have used corpora of primarily fictional literature. The focus on such genres is motivated by the fact that $masse_{MN}$, and in particular $masse_{Q}$, are typical of informal language. To investigate the stylistic value of these words, the newest fiction corpus is compared with corpora from other genres: newspapers, journals (thematically specialized, but not necessarily academic), and laws and official reports. Laws and official reports are very formal genres where informal language is unlikely to be used, while thematic journals are intermediate in formality between laws and reports and fiction. Newspapers are mostly informal. The studied corpora are as follows:

- *Tekstsamlingen* 'The Text Collection' (TxtC), comprising primarily fiction, but also letters and other genres, mostly from the 19th century;[7]
- subcorpora of The Oslo Corpus of tagged Norwegian texts (Bokmål) (OsloK): novels from (a) 1937, (b) 1957, (c) 1977, and (d) laws and Official Norwegian Reports (NOUs) from the period 1981–95;[8]

7 www.dokpro.uio.no/litteratur
8 www.tekstlab.uio.no/norsk/bokmaal

- subcorpora of The Lexicographic Corpus for Norwegian Bokmål (about 1985–2013) (LBK): (a) fictional literature, (b) national, regional, and local newspapers, and (c) journals.[9]

These corpora were searched for tokens of *masse* and *Masse*. The search in the lexicographic fiction corpus was limited to 500 randomly selected hits, while the other searches included all hits in the specified (sub)corpora. The hits were collected in a spreadsheet and categorized semantically and syntactically. First, the tokens were categorized as $masse_Q$, $masse_{MN}$ or other uses of $masse_N$.[10] Second, the tokens of $masse_Q$ and $masse_{MN}$ were categorized according to the type of substance nominal: singular, plural or none (including adverbial uses and cases of an implicit substance nominal).

The quantitative results of the corpus studies are summarized in Tables 1 and 2. While $masse_{MN}$ accounts for less than half the tokens in the oldest texts and $masse_Q$ is absent, together they amount to about 90% in all the later fictional corpora as well as modern newspapers. In modern laws and reports, there are very few cases; the other meanings of $masse_N$ dominate completely. The corpus of journals takes an intermediate position.

	19th c.		1937		1957		1977	
	N	%	N	%	N	%	N	%
en (A) $masse_{MN}$	82	43.9	23	74.2	13	65.0	31	68.9
+ sg.	24	12.8	5	16.1	2	10.0	12	26.7
+ pl.	51	27.3	11	35.5	7	35.0	11	24.4
other	7	3.7	7	22.6	4	20.0	8	17.8
$masse_Q$	–	–	5	16.1	5	25.0	9	20.0
+ sg.	–	–	2	6.5	3	15.0	5	11.1
+ pl.	–	–	3	9.7	1	5.0	4	8.9
other	–	–	–	–	1	5.0	–	–
SUM $masse_{MN+Q}$	82	43.9	28	90.3	18	90.0	40	88.9
Other $masse_N$	105	56.1	3	9.7	2	10.0	5	11.1
SUM total	187	100.0	31	100.0	20	100.0	45	100.0

Table 1: *Masse* in corpora of mostly fiction up to 1977. The labels + sg. and + pl. refer to the number of the following substance nominal. There are no examples of prepositional measure noun constructions.

9 www.hf.uio.no/iln/tjenester/kunnskap/sprak/korpus/skriftspraksk orpus/lbk/

10 The 'rest' category includes cases of *den* (adjective) *masse* 'the (adjective) amount/mass', especially in 19th century texts. This use is not a precursor of $masse_Q$, which is used virtually exclusively in indefinite phrases. Further, it is particularly difficult to differentiate between measure and non-measure use of $masse_N$ in these cases.

There are no examples of $masse_Q$ in the oldest texts, but it has a clear presence in 1937 fiction with about a sixth of the *masse* tokens, growing to more than half in the latest period of fiction (as well as journals) — and more than two thirds in modern newspapers. There is only one example in the modern laws and reports, i.e. less than 1%. While there are more tokens of $masse_{MN}$ than of $masse_Q$ up to 1977, the opposite holds in all the modern corpora except for laws and reports.

As noted above, $masse_{MN}$ is used with both count (plural) and noncount (singular) substance nominals, and $masse_Q$ continues this flexibility. However, there is a tendency towards differentiation in relative numbers. $Masse_{MN}$ clearly prefers plural substance nominals over singulars, and the tendency seems to have grown stronger over time, with plurals almost twice as frequent as singulars. $Masse_Q$ seems to have gone from a weak preference for plural substance nominals in 1937 fiction to a weak preference for singulars in the youngest texts — the difference between the singular and the plural is small, but remarkably similar across genres.

	Fiction		Newspapers		Journals		Laws/reports	
	N	%	N	%	N	%	N	%
en (A) $masse_{MN}$	186	37.2	46	22.5	96	20.0	2	1.9
+ sg.	55	11.0	12	5.9	24	5.0	–	–
+ pl.	102	20.4	*25	12.3	***61	12.7	2	1.9
other	29	5.8	9	4.4	11	2.3	–	–
$masse_Q$	263	52.6	139	68.1	251	52.2	1	0.9
+ sg.	119	23.8	67	32.4	***113	23.5	–	–
+ pl.	107	21.4	**61	29.9	105	21.8	1	0.9
other	37	7.4	11	5.9	33	6.9	–	–
SUM $masse_{MN+Q}$	449	89.8	185	90.7	347	72.1	3	2.8
Other $masse_N$	51	10.2	19	9.3	134	27.9	103	97.2
SUM total	500	100.0	204	100.0	481	100.0	106	100.0

Table 2: *Masse* in modern corpora of different genres. The labels + sg. and + pl. refer to the number of the following substance nominal. *This number includes one prepositional example with *med*. **This number includes one prepositional example with *av* in clefting of the substance nominal, where this preposition is compulsory. ***Each of these numbers includes two prepositional examples with (noncompulsory) *av*.

6 The larger picture: the variability of measure noun constructions

The modern Norwegian juxtapositional measure noun construction stems from an older construction with a genitive-marked substance nominal (e.g. Old Norse *alin vaðmáls* '(an) ell of frieze' with *-s* marking the genitive). Like the prepositional con-

structions, this older construction appears to show that the substance nominal is subordinate to the measure noun. On the other hand: "The structure of juxtapositional pseudopartitives [...] has been what we may call a classic problem: Are such expressions headed by the measure noun or by the substance noun?" (Kinn 2001, p. 2; cf. Diderichsen 1957, p. 241–242; Teleman 1969, p. 22–36; Lødrup 1989, p. 83–86; Delsing 1993, p. 200–223).

Indefinite juxtapositional expressions have no phrase-internal structure showing subordination of one noun to the other. Phrase-external evidence can primarily be found in agreeing adjectival predicates (and, in Nynorsk and some dialects, perfect participles). It is hard to find good evidence from usage, since the combination of indefinite subjects and predicate complement constructions is infrequent. But the available evidence seems to point to a difference between primary and secondary measure nouns. Faarlund et al. (1997, p. 240, 769–70) note that in constructions with a primary measure noun, as exemplified in (30), the substance noun tends to trigger agreement; recall that this is the case for constructions with *masse*$_{MN}$. In my judgement, agreement with the substance noun is the only option in this case, as for other primary measure nouns (of specific number, e.g. *million*; of indefinite number, e.g. *rekke* 'series, number'; of conventional measures, e.g. *liter* 'litre'; and of indefinite quantity, e.g. *masse*).

(30) *En mengde sardiner var råtne/?*råtten*
a.M quantity(M) sardines be.PST rotten.PL/rotten.SG
'A lot of sardines were rotten'

In constructions with a secondary measure noun, as in (31), the measure noun tends to trigger agreement, according to Faarlund et al. According to my intuition, agreement with the substance noun is still the preferred option in (31), although agreement with the measure noun is more acceptable here than in (30).

(31) *En boks sardiner var råtne/?råtten*
a.M tin(M) sardines be.PST rotten.PL/rotten.SG
'A tin of sardines was rotten'

Hankamer and Mikkelsen (2008, p. 326) report that an attempt at collecting acceptability judgements of similar agreement options for Danish produced inconclusive results, which made them leave out such data; arguably, the vacillation may be regarded as evidence for variable structure. In light of their origin in genitival constructions, juxtapositional constructions appear partly to have undergone reanalysis, i.e. from (simplified) [N [N]] to [[N] N], and the reanalysed structure seems to be more strongly conventionalized for primary than for secondary measure nouns. Vacillation in agreement may then be accounted for as due to variation between the old and the new structure

(see e.g. Delsing 1993). The development from [N [N]] to [[N] N] may be regarded as an indication of ongoing grammaticalization of the measure noun (see Section 8).

At first sight, prepositional expressions appear to have the (simplified) structure [N [P [N]]]. But such constructions, too, exhibit vacillating agreement properties, see (32) and (33).

(32) *En mengde med sardiner var råtne/?*råtten*
a.M quantity(M) with sardines be.PST rotten.PL/rotten.SG
'A lot of sardines were rotten'

(33) *En boks med sardiner var råtne/råtten*
a.M tin(M) with sardines be.PST rotten.PL/rotten.SG
'A tin of sardines was rotten'

Recall from Section 3 that prepositional constructions with *en masse med* exhibit substance noun agreement. Agreement with the substance noun and vacillating agreement is found also in English, viz. in the agreement inflection of verbs in the present tense (plus *was*/*were*), e.g. as in (34)–(36) (cf. Langacker 1991, p. 88–89). Similar properties have been documented for Spanish prepositional measure noun constructions, e.g. (37), where the finite verb *acercan* agrees with *personas* rather than with *aluvión* (Delbecque and Verveckken 2014, p. 94–95).

(34) A lot of students were in the room

(35) A bunch of carrots was in the sink

(36) A bunch of students were in the room

(37) *Un aluvión de personas se le acercan*
a flood of persons REFL him approach
'A flood of persons approach him'

The adjectival or verbal agreement with the (apparently subordinate) substance noun in the apparent structure [N [P [N]]] is not straightforwardly accounted for. It might be regarded as semantic agreement, i.e. agreement that disregards the syntactic structure. Such an account could be extended to juxtapositional measure noun constructions: It would then not be necessary to assume that reanalysis had taken place there; the structure would be [N [N]] regardless of agreement properties. This seems to be the view of Faarlund et al. (1997, p. 769–770).

However, several researchers on English and Spanish have argued that substance noun agreement is evidence that syntactic reanalysis has taken place even in prepositional structures (e.g. Delbecque and Verveckken 2014; Traugott and Trousdale 2013). That is, there has been a change from [N [P [N]]] to something like [[N P] N], e.g. [[*a*

bunch of] *students*]. A different structure, [[N] [P N]], was proposed for Norwegian by Kinn (2001, p. 216–220), where the substance noun is the head and the preposition has become a head marker. Both analyses would account for external agreement properties, but the internal structure of the constructions is in both cases somewhat obscure.

The exact analyses of constructions headed by the substance noun will not be discussed in further detail here, since the focus is on structures where a former measure noun has become a quantifier (in terms of its word class, not just its function). What matters is that there does appear to be a change going on which switches head status from measure noun to substance noun, and which, in prepositional constructions, renders the status of the preposition unclear. This change is evidently a reanalysis whose semantic motivation is strong enough to override the quite transparent previous [N [P [N]]] structure.

If the agreement of constituents external to the measure noun construction had been the only evidence for the restructuring, one might have argued that we are dealing with purely semantic agreement, and that the measure noun construction is always headed by the measure noun. However, in Norwegian there is also evidence from internal structure that there is more going on.

Not only adjectival predicate complements but also a nominal-internal plural determiner (definite article, demonstrative) may in some cases agree with a plural substance noun — 'across' the measure noun and (if present) a preposition. To demonstrate this, the Norwegian opposition between single and double definiteness must first be presented.

The term 'single definiteness' is used primarily about nominal constructions with a definite article followed by a quantifier and/or an adjective and an indefinite noun. This is mostly a conservative feature of written Bokmål, but is nevertheless common when followed by certain restrictive modifiers, especially restrictive relative clauses. An example is given in (38), where *spørsmål* is indefinite. The article *de* and *spørsmål* agree in number, but disagree in definiteness.

(38) *de* *mange* *vanskelige* *spørsmål* *(som* *styret* *stiller)*
the.PL many.PL difficult.PL questions that the.board asks
'the many questions (that the board is asking)'

More commonly, the noun is in the definite form, yielding 'double definiteness'. This is exemplified in (39), where *spørsmålene* is definite. The article *de* and the noun *spørsmålene* agree both in number and in definiteness.

(39) *de* *mange* *vanskelige* *spørsmålene* *(som* *styret* *stiller)*
the.PL many.PL difficult.PL the.questions that the.board asks
'the many questions (that the board is asking)'

Examples (38) and (39) involve the quantifier *mange* 'many' modifying the substance noun with respect to its quantity. The distinction between single and double definiteness is also found in measure noun constructions. With two nouns involved, there are in principle two candidates for definiteness inflection in double definiteness and for the definite article to agree with.

Numeral nouns are the class of measure nouns apparently most prone to develop into quantifiers (see Section 8). They exhibit several constructional patterns and will serve to illustrate some essential points below. In single definiteness, the form of the nouns provides no clue to which one is the head, since both are indefinite, as shown for juxtapositional and prepositional measure noun constructions, respectively, in (40) and (41):

(40) *alle de millioner mennesker som følger med på fotball*
all.PL the.PL millions human.beings that follow with on football
'all the millions of people that follow football'

(41) *alle de millioner av mennesker som trenger hjelp*
all.PL the.PL millions of human.beings that need help
'all the millions of people that need help'

In double definiteness, the numeral noun may be definite and the substance noun indefinite, showing the headhood of the former, exemplified for juxtapositional and prepositional measure noun constructions, respectively, in (42) and (43):

(42) *alle de millionene mennesker som ønsker å se Ham*
all.PL the.PL the.millions human.beings that wish to see Him
'all the millions of people that wish to see Him'

(43) *alle de millionene med mennesker som verken kan lese*
all.PL the.PL the.millions with human.beings that neither can read
eller skrive
nor write
'all the millions of people that can neither read nor write'

However, it is probably more common to have the numeral noun in the indefinite and the substance noun in the definite form, thus with the latter as head, as shown for juxtapositional and even for prepositional measure noun constructions in (44) and (45), respectively:

(44) *alle de millioner menneskene som er preget etter*
all.PL the.PL millions the.human.beings that are marked after
kommunismen
the.communism
'all the millions of people that are marked as a result of communism'

(45) *alle de millioner av menneskene som er på flukt fra*
all.PL the.PL millions of the.human.beings that are on flight from
denne meningsløse krigen
this meaningless the.war
'all the millions of people that are on the run from this meaningless war'

Constructions with numeral nouns allow an indefinite measure noun in the singular to appear between a plural article and a definite plural substance noun, as illustrated in (46) and (47):

(46) *Det hersker stor spenning blant de ett tusen*
it rules great excitement among the.PL one.N thousand(N)
bøndene i Fjellregionen
the.farmers in the.Mountain.Region
'There is much nervous anticipation among the farmers of the Mountain Region'

(47) *… bør i alle fall to av de en million eggene i*
… ought in all cases two of the.PL one.M million(M) eggs in
denne rogna vokse opp
this roe grow up
'should at least two of the one million eggs in this roe grow up'

These data confirm the rather vague indications from agreement data and indefinite measure noun constructions: The substance noun can be head, and headhood status may even override the prepositional marking.

The situation described for numeral nouns is far from common to all definite measure noun constructions. Most juxtapositional expressions show the measure noun to be superordinate, e.g. (48) in which the determiner *de* agrees with the measure noun *literne*). An expression like (49), with singular *den* agreeing with the substance noun *vinen*, is quite ill-formed. Prepositional expressions typically also have a structure indicating that the measure noun is the head, e.g. (50) with agreement between determiner and measure noun.

(48) *de tre literne vin*
the.PL three the.litres wine
'the three litres of wine'

(49) **den tre liter vinen*
the.SG three litres the.wine

(50) *de tre literne med vin*
the.PL three the.litres with wine
'the three litres of wine'

Apparently, double definiteness involving a substance noun requires that it and the article (or demonstrative) both be in the plural, and the measure noun must — if it is not a numeral noun — be in the abundance plural. Such expressions are not very common, and not everybody finds them quite acceptable. But it is my intuition — built on two decades of interest in abundance plurals — that they are becoming steadily more conventional; (51)–(55) provide illustration and give an impression of the kind of structure we are dealing with.

(51) *Alle forgreiningene og de tusenvis av lungeblærene*
all the.branchings and the.PL thousand.ABU of the.alveoli
renses og holdes åpne
are.cleaned and are.held open
'All the branches and the thousands of alveoli are kept clean and open'

(52) *men av alle de tonnevis av skytespillene på markedet er*
but of all.PL the.PL ton.ABU of the.shooting.games on the.market is
det veldig lite som genuint interesserer meg
it very little that genuinely interests me
'but among all the tons of shooting games on the market, there is very little that genuinely interests me'

(53) *man må bruke traktor på de milevis med grusveiene opp*
one must use tractor on the.PL mile.ABU with the.gravel.paths up
til bondelandet
to the.farm.land
'one has to use a tractor on the miles of gravel paths up to the farm land'

(54) *Det eneste problemet vil være desentraliseringen og alle de*
the only the.problem will be the.decentralization and all.PL the.PL
drøssevis med nettverkene
ton.ABU with the.networks
'The only problem will be decentralization and all the tons of networks'

(55) *alle de massevis av produktene som inneholder billige raffinerte*
all.PL the.PL lot.ABU of the.products that contain cheap refined
planteoljer
plant.oils
'all the tons of products that contain cheap refined plant oils'

It seems quite clear in these examples that there is agreement between the definite plural article *de* and the definite plural substance noun, in spite of the intervening preposition.

The examples in (51)–(55) all have double definiteness. Single definiteness is quite common, provided that there is a restrictive modifier, typically a relative clause, as in (56):

(56) *Vi har jo kun besøkt et fåtall av alle de*
we have of.course only visited a minority of all.PL the.PL
hundrevis av campingplasser som finnes i vårt langstrakte
hundred.ABU of camp.sites that exist in our long-stretched
land
country
'Of course, we've only visited a small minority of all the hundreds of camp sites that there are in our long-stretched country'

If there is no restrictive modifier (e.g. if the relative clause of (56) were left out), the result is stylistically clearly marked (conservative). This shows that it is the substance noun that partakes in the single vs. double definiteness distinction and is the head of the measure noun construction.

This rather long discussion has demonstrated that some measure nouns are subordinate to the substance noun of measure noun constructions. Importantly, as shown in Section 3, this holds for *masse*$_{MN}$.

7 The larger picture: changes in measure noun constructions

To gain a better understanding of the synchrony of measure noun constructions, it is useful to start with constructions that may be assumed to precede them diachronically. Discussing English measure noun constructions, Langacker (1991, p. 88) notes that some of the measure nouns (i.e. those here called secondary measure nouns) "have an interpretation in which they designate a physical spatially-continuous entity that either serves as the container for some portion of a mass (*bucket, cup,* […]), or else is constituted of some such portion (*bunch, pile,* […])". Norwegian measure noun constructions with *med* 'with' and *av* 'of' illustrate well the two conceptions of quantity described by Langacker. The use of *med* clearly evokes the conceptual relation between a container and its content, while the use of *av* evokes the relation between an object and its constitutive material (see Kinn 2001, p. 174–179).[11] But neither of these conceptions are inherently quantifying. Nonquantifying uses illustrating this may be *ei lommebok med 300 kroner* 'a wallet with 300 kroner' and *ei jakke av skinn* 'a jacket (made) of leather'. In such cases, the syntactic structure is unambiguous (simplified: [N [P [N]]]). The relations denoted by the prepositions are understood literally, and the

11 Kinn (2001, p. 172–174) argues that the use of *med* in Norwegian has an additional relevant meaning that motivates an observed stronger preference for it in constructions of length and time, namely that of the relation between something accompanied and its accompaniment.

nouns involved are not coextensive.[12] The nominals may appropriately refer to wallets and jackets, but not to kroner and leather.

In measure nouns constructions, however, the nouns are coextensive, or as Kinn (2001, p. 5–6) says, they are weakly coreferential. In that work, it is regarded as a defining characteristic of measure noun constructions that the nominals refer to the same entity but categorize it in different ways. Thus, in (57), the 'lot' and the students are the same entity. In (58), the litres and the wine are the same. The measure noun refers to the mass by specifying its quantity,[13] while the substance noun provides qualitative information. The weakness of the coreferentiality lies in the difference in semantic substructures of the nouns. For instance, in (59), the collective of kilos and the collective of potatoes are the same whole entity, but the individual kilos and the individual potatoes are different entities. Verveckken (2015), dealing with Spanish, analyses measure noun constructions in a very similar way to Kinn (2001).

(57) *en masse studenter*
a.M lot(M) students
'a lot of students'

(58) *to liter (med) vin*
two litres (with) wine
'two litres of wine'

(59) *fire kilo poteter*
four kilos potatoes
'four kilos of potatoes'

The coreferentiality of both nouns is evident in Norwegian pairs like (60) and the closely synonymous (61). The prepositions *med* and *i* are converses, the former relating a container to a content and the latter relating a content to a container. But here, the containment is metaphorical; container and content are the same.

(60) *litervis med vin*
litre.ABU with wine
'litres of wine'

(61) *vin i litervis*
wine in litre.ABU
'litres of wine'

12 Assuming the jacket has a lining etc. in other materials.
13 Secondary measure nouns also contribute some qualitative information.

The quantity of the substance is contributed more or less clearly by the measure noun. Secondary measure nouns do not specify an accurate quantity, but they tend to have a typical size associated with them, and this is how they come to be able to serve a quantifying function. Further, "these size implications can be foregrounded through pragmatic enrichment, to the detriment of the lexical meaning" (Brems 2011, p. 108–109). Some English measure nouns, like *bunch*, which have until recently been secondary measure nouns with a fairly clear qualitative meaning (e.g. *a bunch of carrots*), have developed a more general quantitative meaning, i.e. have become primary measure nouns (e.g. *a bunch of students, a bunch of rubbish*). The further this development goes, the more quantifier-like the measure noun becomes, and the more head-like the substance noun becomes.

The observed facts have explanatory power in relation to diachrony. The coextensiveness of the nouns explains why it matters little in terms of reference whether one or the other noun heads the referring expression. A reversal in head status between measure noun and substance noun corresponds to a subtle figure–ground reversal in the conceptual semantics of the measure noun construction — a metonymic shift. Given that the nouns are coextensive, the preposition in prepositional measure noun constructions (*med* or *av* in Norwegian, *of* in English) is of little referential importance. This explains why the clear syntactic hierarchy of such structures may be overridden in a semantically-based reanalysis, promoting the substance noun to head status and demoting the measure noun.

As the data and discussion above have shown, *masse*$_{MN}$ is among the demoted primary measure nouns in constructions involving the indefinite singular *en masse* and partly the abundance plural *massevis*.

8 From primary measure noun to quantifier

Constructions with primary measure nouns that have become subordinate to the substance noun in some cases continue into a development where the measure noun loses noun properties and instead acquires the modifier properties of a quantifier. One measure noun that has wholly undergone such a development is the predecessor of *ti* 'ten' (now only a quantifier; compare modern Norwegian *seksti* '60' to Old Norse *sex tigir* [six tens]). The larger numeral nouns *hundre* 'hundred' and *tusen* 'thousand' exhibit some uses where they may be regarded as quantifiers, and so does *par* 'couple' (Kinn 2000). *Masse* is perhaps the youngest example.

The developments described above for Norwegian *masse* exhibit a number of characteristics of grammaticalization (see e.g. Lehmann 2015). When *masse*$_{N}$ develops the meaning variant of *masse*$_{MN}$, this is a case of desemanticization or bleaching. It is also a case of paradigmatization when the noun enters into the paradigm of measure nouns (which is rather large, but very much smaller than the paradigm of nouns in general). When *en masse* and *masse* come to be used as adverbial quantifiers and quantify

over predicates in addition to nominal entities (as in (3) and (6) above), this is context expansion, which according to some theorists (e.g. Himmelmann 2004) is typical of grammaticalization. Paradigmatization continues with the development of a quantifier, since the class of quantifiers is rather limited compared to that of measure nouns. This downgrading change involves loss of nominal properties (i.e. decategorialization), namely gender, inflection for number and definiteness. But it also involves gain of the adjectival property of gradability (accepting degree modifiers). The developments have led to divergence (the expression *masse* belongs to different categories) and layering (*masse*$_Q$ is a young member of a paradigm together with e.g. older *mye*$_Q$ 'much' and *mange*$_Q$ 'many').

Although the development from *masse*$_{MN}$ to *masse*$_Q$ may be regarded as a natural diachronic change, it also illustrates the piecemeal nature of language change. The development from a meaning of '(shapeless) matter, substance' to a purely quantitative meaning and further from noun to quantifier appears to have started not many generations after it was borrowed. The measure noun *mengde* 'lot, quantity' is similar in meaning and much older (derived from *mang(e)* 'many') but does not appear to be developing a quantifier variant. The different fates of *masse* and *mengde* may however not be accidental. Although both *en masse* and *en mengde* mean 'a lot', a difference comes out if we look at their use with adjectives. *Mengde*$_{MN}$ is modified by adjectives of both large and small size (*en stor mengde* 'a large number/amount', *en liten mengde* 'a small number/amount'). *Masse*$_{MN}$ is infrequently used with adjectives of size (*stor masse* and *liten masse* typically refer to great and small mass in the physics meaning). Instead, it tends to be used with intensifying adjectives (*en hel masse* 'a whole lot', *en helvetes masse* 'a helluva lot'), which only go upwards. Thus, while the size meaning of *mengde*$_{MN}$ is manipulable in both directions, *masse*$_{MN}$ normally indicates only large amount. In that sense, the inherent meaning of *masse*$_{MN}$ makes it a better candidate for quantifier-hood than *mengde*$_{MN}$.

9 Conclusion

The Norwegian quantifier *masse* 'a lot, lots, many, much' has developed from the measure noun *masse* 'matter, mass, large amount'. The development must probably have begun in the late 19th century, and the use of *masse* as a quantifier seems to have become conventionalized in informal language during the first few decades of the 20th century. In contemporary Norwegian, it is quite frequent, but it is still limited to informal language and hardly found in more formal text types such as laws and governmental reports. The development of a quantifier from a measure noun has been shown to be facilitated by the inherent variability of measure noun constructions, where semantically motivated reanalyses demote measure nouns from heads to quantifying modifiers. Such demotion may be regarded as a first step towards grammaticalization from noun to quantifier.

Acknowledgements

I wish to thank the two anonymous referees, whose suggestions have helped me to improve this text.

Appendix: Sources of examples

[1-6] LBK, fiction
[7–8] Author's examples.
[9] *Norske Intelligenssedler*, 1773, from nb.no
[10] *Den Norske Rigstidende*, 1819, from nb.no
[11] LBK, newspapers
[12] LBK, textbooks
[13] LBK, newspapers
[14] LBK, journals
[15–16] Norsk Aviskorpus (NAK), avis.uib.no
[17] George Manville Fenn, *Et dobbelt Problem* (translated from English, anonymous translator), from nb.no
[18] George de Horne Vaizey, *Huset ved veien* (translated from English by Ingeborg von der Lippe Konow), from nb.no
[19] Adam Hiorth, "Barnehjælpsdag" in *Det Norske studentersamfunds viser og sange gjennem hundrede aar: 1813–1913*, from nb.no
[20] Julli Wiborg, *Ragna*, from nb.no
[21] George de Horne Vaizey, *Darsie* (translated from English by Ingeborg von der Lippe Konow), from nb.no. Found by one of the anonymous reviewers.
[22] John Bergh, *Den vidunderlige globus* (translated from Swedish by G. Emil Thomassen), from nb.no
[23] *Norsk riksmålsordbok* (1937-1957), vol. 2, part 1
[24] www.annicaolsson.se
[25] esmark.dk
[26] org.ntnu.no
[27] tormodsgate8.weebly.com
[28] TxtC
[29] LBK, fiction
[30–33] Faarlund et al. (1997, p. 240). Author's grammaticality judgements.
[34–36] Langacker (1991).
[37] Delbecque and Verveckken (2014).
[38–39] Author's examples.
[40] shop.flammeforlag.no
[41] www.adventist.no
[42] www.verdidebatt.no
[43] paeliassen.no
[44] unitedforumet.no
[45] nettavisen.no
[46] www.retten.no
[47] fiskeribladet.no
[48–50] Author's examples.
[51] LBK
[52] www.gamereactor.no
[53] vgd.no
[54] www.diskusjon.no

[55] www.helhetligliv.no
[56] bobilverden.no
[57–61] Author's examples.

References

Bokmålsordboka (2005). 3rd ed. Oslo: Kunnskapsforlaget.

Brems, Lieselotte (2011). *Layering of Size and Type Noun Constructions in English*. Berlin: De Gruyter.

Clerck, Bernard De and Lieselotte Brems (2015). "Size nouns matter: a closer look at *mass(es)* of and extended uses of SNs". In: *Language Sciences* 53, pp. 160–176.

Delbecque, Nicole and Katrien Dora Verveckken (2014). "Conceptually driven analogy in the grammaticalization of Spanish binominal quantifiers". In: *Linguistics* 52, pp. 637–684.

Delsing, Lars-Olof (1993). *The Internal Structure of Noun Phrases in the Scandinavian Languages*. Lund: Dept. of Scandinavian Languages, University of Lund.

Diderichsen, Paul (1957). *Elementær Dansk Grammatik*. 3rd ed. Copenhagen: Gyldendal.

Enger, Hans-Olav and Philipp Conzett (2016). "Morfologi". In: *Norsk språkhistorie*. Ed. by Helge Sandøy. Vol. 1: Mønster. Universitetsforlaget, pp. 213–315.

Faarlund, Jan Terje, Svein Lie, and Kjell Ivar Vannebo (1997). *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Hankamer, Jorge and Lise Mikkelsen (2008). "Definiteness marking and the structure of Danish pseudopartitives". In: *Journal of Linguistics* 44, pp. 317–346.

Himmelmann, Nikolaus P. (2004). "Lexicalization and grammaticalization: Opposite or orthogonal?" In: *What Makes Grammaticalization: A Look from its Components and its Fringes*. Ed. by Walter Bisang, Nikolaus P. Himmelmann, and Björn Wiemer. Berlin: Mouton de Gruyter, pp. 21–42.

Kinn, Torodd (2000). "Eit par tankar om *par*". In: *Menneske, språk og fellesskap. Festskrift til Kirsti Koch Christensen på 60-årsdagen 1. desember 2000*. Ed. by Øivin Andersen, Kjersti Fløttum, and Torodd Kinn. Oslo: Novus, pp. 135–140.

– (2001). "Pseudopartitives in Norwegian". PhD thesis. University of Bergen.

– (2004). "*Stribevis af eksperter, drøssevis av sjansar, hinkvis med kaffe*: An emerging plural of abundance in the Scandinavian languages". In: *Proceedings of the 20th Scandinavian Conference of Linguistics, Helsinki, January 7-9, 2004*. Ed. by Fred Karlsson. Helsinki: Dept. of General Linguistics, University of Helsinki.

– (2005). "Ord på *-vis* i moderne norsk: samansetningar, avleiingar — og bøyingsformer?" In: *Maal og Minne* 2005, pp. 45–78.

Koptjevskaja-Tamm, Maria (2001). ""A piece of the cake" and "a cup of tea": Partitive and pseudo-partitive nominal constructions in the Circum-Baltic languages". In: *Circum-Baltic Languages*. Ed. by Östen Dahl and Maria Koptjevskaja-Tamm. Vol. 2: Grammar and Typology. Amsterdam: Benjamins, pp. 523–568.

Langacker, Ronald W. (1991). *Foundations of Cognitive Grammar*. Vol. 2: Descriptive Application. Stanford, CA: Stanford University Press.

Lehmann, Christian (2015). *Thoughts on Grammaticalization*. 3rd ed. Berlin: Language Science Press.

Lødrup, Helge (1989). *Norske hypotagmer*. Oslo: Novus.

Norsk riksmålsordbok (1937-1957). Oslo: Aschehoug.

Teleman, Ulf (1969). *Definita och indefinita attribut i nusvenskan*. Lund: Studentlitteratur.

Traugott, Elizabeth Closs and Graeme Trousdale (2013). *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.

Verveckken, Katrien Dora (2015). *Binominal Quantifiers in Spanish*. Berlin: De Gruyter.

Reflexive sentences with *la* 'let' in Norwegian — active or passive?

Helge Lødrup

Abstract. This article discusses Norwegian sentences such as *Helge lar seg ikke stoppe* 'Helge lets REFL not stop'. The second verb raises a difficult question: It does not have passive morphology, but it seems to share properties with passive verbs. This problem has been discussed for corresponding constructions in e.g. German and French. I focus on the Norwegian data, and argue that it is necessary to consider this kind of sentence to be passive. I also discuss how to implement this view within an LFG conception of complex predicates.

1 Introduction

What is a passive verb? The question might seem trivial, and it is not often asked. However, Dyvik (1980) did ask, both for Old Norse and for general grammatical theory. He stressed the structuralist principle of solidarity between content and expression. To assume a passive, there must be a certain content — the well known change in the relation between thematic roles and syntactic functions — combined with an identifiable expression (Dyvik 1980, p. 91). In practice, this means that there must be some kind of morphological marking. This requirement has also been stressed by others, e.g. Haspelmath (1990).

A possible problem for this requirement is represented by some causative and causative-like constructions, with verbs such as German *lassen* 'let' or French *faire* 'make'. A German example is (1), from Comrie (1976, p. 271). The second verb has the active form. Even so, it seems to share properties with passive verbs: it has an agent phrase, and its logical object could be argued to be in a subject position.

(1) *Er liess den Brief von seinem Sohn abtippen.*
he let the-ACC letter by his-DAT son type
'He made his son type the letter.'

This question has been discussed many times for various languages (see e.g. Comrie 1976, pp. 271–75; Haspelmath 1990, pp. 46–49; Pitteroff 2014). The facts are complicated, and they are not identical from language to language. Comrie may have been

The very model of a modern linguist edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1328.

right when he wrote that the question of voice depends upon detailed study of the individual languages (Comrie 1976, pp. 272).

The question of voice also arises in sentences that have a reflexive pronoun with the first verb, such as the Norwegian (2) with the verb *la* 'let'.

(2) *Helge lar seg ikke stoppe av hindringer.*
Helge lets REFL not stop by obstacles
'Helge cannot be stopped by obstacles.'

Similar sentences have been discussed for Germanic and Romance languages (see e.g. Pitteroff 2014 for German; Labelle 2013 for French). In this article, I will discuss whether Norwegian sentences such as (2) should be considered passive, a question which has not been raised in the Scandinavian literature (Taraldsen 1983; Taraldsen 1991; Vikner 1987). Section 2 shows how *la* 'let' in reflexive sentences is different from other uses of this verb. Section 3 discusses properties of these sentences that could provide arguments for or against a passive analysis. Sections 4 and 5 argue that the point of departure for an analysis must be the theory of complex predicates, and give an account based upon Lexical Functional Grammar (LFG). Section 6 discusses cases where the second verb has passive morphology.

2 The verb *la* 'let'

The Norwegian verb *la* 'let' (henceforth LA) can take a verbal complement in sentences such as (3)–(6).[1]

(3) *Vi lot vaktene løslate fangene.*
we let guards-DEF release prisoners-DEF
'We let the guards release the prisoners.'

(4) *Vi lot løslate fangene.*
we let release prisoners-DEF
'We let the prisoners be released.'

(5) *Vi lot fangene løslate.*
we let prisoners-DEF release
'We let the prisoners be released.'

(6) *Fanger lar seg gjerne løslate.*
prisoners let REFL gladly release
'Prisoners are happy to be released.'

1 Examples (4)–(5) are from Taraldsen (1983, p. 201).

In (3), the logical subject of the second verb is raised to be the object of LA. In (4) and (5), the second verb has a realized logical object, but no realized logical subject. These sentences will be referred to as 'prisoner sentences'. The standard claim in the literature is that Norwegian has two options for word order in prisoner sentences, with the logical object following or preceding the embedded verb. This claim can be found from Falk and Torp (1900, p. 200) to Taraldsen (1983, p. 201) and Taraldsen (1991). Norwegian has been compared to Danish, which has the first word order only, and Swedish, which has the second only (e.g. Taraldsen 1991). However, in current colloquial Norwegian, prisoner sentences are archaic, especially the type in (5), (see e.g. Faarlund et al. 1997, p. 1009). Even if examples can be found in texts, it would be only a mild idealization to say that prisoner sentences no longer exist as a productive option.

While prisoner sentences are archaic, sentences such as (6), with a reflexive following LA are perfectly normal (Taraldsen 1983, p. 225; Faarlund et al. 1997, p. 1009). This type of sentence will be referred to as reflexive LA sentences.

The relation between the uses of LA in (3)–(6) raises some questions. Examples (5) and (6) might look rather similar from a syntactic point of view. An important difference between the sentence types is that the noun phrase following LA realizes the logical object of the second verb, while the reflexive does not. The logical object of the second verb in sentences such as (6) is often realized as the subject of LA (see Taraldsen 1983, p. 233 and Vikner 1987, pp. 271–72 on Norwegian and Danish). Examples (7)–(8) from Taraldsen (1983, p. 233) illustrate how the syntactic and semantic properties of the subject of LA are constrained by the second verb. The clausal subject in (7)–(8) only gives meaning when the second verb is of a type that can take a clausal object, as in (7).

(7) *At jorden er flat lar seg neppe hevde idag.*
that earth-DEF is flat lets REFL hardly claim today
'That the earth is flat can hardly be claimed today.'

(8) *#At jorden er flat lar seg neppe hjelpe over gaten.*
that earth-DEF is flat lets REFL hardly help across street-DEF
'That the earth is flat can hardly be helped across the street.'

The meanings of prisoner sentences are rather different from those of reflexive LA sentences. In the prisoner sentences, the subject is a causer. In reflexive LA sentences, on the other hand, the meaning is not causative. In many cases, the predicate denotes something that happens to the subject, as in (9), or a disposition that the subject has, as in (10).

(9) *Ola lot seg behandle.*
Ola let REFL treat
'Ola was treated.'

(10) *Sykdommen lar seg ikke behandle.*
disease-DEF lets REFL not treat
'The disease is untreatable.'

In reflexive LA sentences, a human subject can be implied to have some control over the event, at least by not opposing it. This fact might seem to stand in the way of a passive analysis, but this is not the case. Furthermore, the subject can be implied to have some control in the regular periphrastic passive (see e.g. Engdahl 2006, pp. 32–34). For example, the periphrastic passive has an imperative, as in (11) (adapted from Engdahl 2006, p. 33), as opposed to the morphological passive. Keenan and Dryer (2007, p. 340) say that "distinct passives in a language are likely to differ semantically with respect to aspect and/or degree of subject affectedness …"

(11) *Ikke bli ranet i Chicago!*
not become robbed in Chicago
'Don't get robbed in Chicago!'

In the literature on German, a traditional idea is that the second verb is passive both in reflexive LA sentences and in prisoner sentences with a preposed logical object (see e.g. Reis 1973; Pitteroff 2014). For Norwegian, Åfarli and Eide (2003, pp. 220–22) claim that prisoner sentences are passive (see also Platzack 1986 on Swedish). I will not discuss prisoner sentences any further, for two reasons. First, the idea of prisoner sentences being passive only gives meaning if the logical object of the second verb is its structural subject. It is not clear, however, that it is not its structural object (see e.g. Gunkel 1999 on German). Second, it is very difficult to argue for or against analyses of prisoner sentences in Norwegian, given their marginal status.

I will first give an overview of facts that seem to indicate that passive voice is in some way involved in reflexive LA sentences in Norwegian. Relevant phenomena concern subject choice, the behavior of the external argument of the second verb, and exceptions to the passive. These kind of phenomena have been discussed for German and other languages (see e.g. Pitteroff 2014 and references there). The Norwegian facts are not identical, but the differences between the languages will not be focused on here.

Most example sentences in the following are from the World Wide Web, found either by googling or by searching the NoWaC-corpus (Norwegian Web as Corpus). Some of them have been edited lightly.

3 A comparison to regular passives

In reflexive LA sentences such as (6), the logical object of the second verb is realized as the subject of LA. There are also other options for choosing a subject in these sentences. These options will now be discussed and compared to those of regular passives.

Impersonals. All regular Norwegian passives have an impersonal version with an expletive subject (see e.g. Åfarli 1992, p. 20). Reflexive LA sentences can also be impersonal. Two examples are (12) with a presentational focus construction (see Taraldsen 1983, p. 231), and (13) with an unergative verb.

(12) *Det lar seg skaffe dokumentasjon.*
EXPL lets REFL provide documentation
'Documentation can be provided.'

(13) *så lenge det lar seg trene på kunstgresset*
as long EXPL lets REFL practice on astroturf-DEF
'as long as we can practice on the astroturf'

Non-thematic subjects. An important fact is that the subject of LA can correspond to an argument that does not get a thematic role from the second verb. In (14)–(15), the derived subject corresponds to the object of the unergative second verb. This argument is also the subject of a resultative predicate. It does not get a thematic role from the second verb, only from the resultative (*bort* 'away' and *flat* 'flat').

(14) *overflødig fett som ikke lar seg trene bort*
excess fat that not lets REFL exercise away
'excess fat that cannot be removed by exercising'

(15) *Ingen skulle la seg tråkke flate.*
nobody should let REFL step flat
'Nobody should let anyone squeeze themselves.'

Sentences such as (14)–(15) give important arguments for a passive analysis. With middles and unaccusatives, a derived subject must be an argument that gets a thematic role from the verb (Keyser and Roeper 1984, p. 409; Levin and Hovav 1995, pp. 42–48). With passives, on the other hand, there is no such requirement (compare *Fettet ble trent bort* 'The fat was exercised away').

Benefactives. Another difference from middles and unaccusatives (Baker 1993) is that the derived subject is not limited to the theme argument. When the second verb is ditransitive, its benefactive argument is usually realized as the subject of LA, as in (16).

(16) *Mussolini lot seg overrekke et sverd.*
Mussolini let REFL present a sword
'Mussolini was presented with a sword.'

In regular Norwegian passives of ditransitives, the subject can correspond to either the theme or the benefactive argument (even if the latter option seems to be more common). In reflexive LA sentences, however, theme subjects are marginal, cf. (17).

(17) *??Dette sverdet lot seg ikke overrekke Mussolini.*
this sword let REFL not present Mussolini
'This sword could not be presented to Mussolini.'

This is a difference between regular passives and reflexive LA sentences. However, Herslund (1986) and Vikner (1987, pp. 274–277) show that double objects in causatives and causative-like sentences behave in ways that are not understood. For example, when a sentence such as (12) is acceptable, it is not easy to see why the corresponding sentence with two objects in (18) is not.

(18) **Det lar seg skaffe ham dokumentasjon.*
EXPL lets REFL provide him documentation
'Documentation can be provided for him.' [intended meaning]

Pseudopassives. An option that is very limited in the world's languages is the pseudopassive, in which the passive subject corresponds to the object of a preposition. No language seems to have a corresponding option with unaccusatives or middles, as has sometimes been pointed out (e.g. Drummond and Kush 2015, p. 458). Norwegian is a language that has pseudopassives, and reflexive LA sentences seem to allow this option to the same extent. Examples are (19) and (20).

(19) *et arbeidsliv der arbeideren lar seg bestemme over*
an economic.life there worker-DEF lets REFL decide over
'an economic life where the worker is controlled'

(20) *reir som ikke lar seg hakke hull på av hakkespetten*
nests which not let REFL peck holes in by woodpecker
'nests which the woodpecker cannot drill holes in'

We see, then, that the options for choosing a subject in reflexive LA sentences are strikingly similar to those in regular passives, with a slight complication for ditransitive verbs. We will now compare two other properties of the passive, namely its exceptions and the behavior of its demoted argument.

Exceptions to the passive. Some verbs cannot passivize. If reflexive LA sentences are passive, these verbs would be expected not to occur as second verbs. This seems to be what we find. Example (21) has an unaccusative verb, while (22)–(23) have verbs whose meanings make them impossible to passivize "in the majority of languages" (Siewierska 1984, p. 189), including Norwegian.

(21) **Det lar seg forsvinne i skogen.*
EXPL lets REFL disappear in woods-DEF
'One can disappear in the woods.' [intended meaning]

(22) **Penger lar seg aldri mangle på universitetet.*
money lets REFL not lack at university-DEF
'One is never short of money at the university.' [intended meaning]

(23) **Elvis lar seg vanskelig ligne.*
Elvis lets REFL hardly resemble
'One can hardly resemble Elvis.' [intended meaning]

There are also language-specific exceptions to the passive. Norwegian does not allow verbs ending in *-s* to passivize, such as *synes* 'think'. The verbs *skylde* 'owe' and *slippe* 'avoid' are idiosyncratic exceptions (Lødrup 2000). It is striking that even these restrictions on the passive seem to be reflected in reflexive LA sentences, as shown in (24)–(26). Verbs with related meanings such as e.g. *tenke* 'think', *avse* 'spare' and *unngå* 'avoid' can be used both in the regular passive and in reflexive LA sentences.

(24) **At filmen var god lar seg vanskelig synes.*
that movie-DEF was good lets REFL hardly think
'One can hardly think that the movie was good.' [intended meaning]

(25) **Hvor mange penger lar seg skylde av et EU-land?*
how many money lets REFL owe by an EU-country
'How much money can an EU country owe? [intended meaning]'

(26) **Rengjøringen lar seg aldri slippe.*
cleaning-DEF lets REFL never avoid
'One can never avoid the cleaning.' [intended meaning]

It is difficult to find clear counterexamples to the generalization that verbs that cannot be passivized do not occur in reflexive LA sentences. A possible case is (27), with the verb *interessere* 'interest'. In my intuition, this verb has no regular passive. The morphological passive can be found in texts, however.

(27) *Jeg håper jo at noen vil la seg interessere.*
I hope you.know that somebody will let REFL interest
'I hope that somebody will be interested, you know.'

The demoted argument. Passives have an implicit external argument, which can be realized as an agent phrase. In some cases, the implicit external argument can be a controller of PRO. Reflexive LA sentences allow an agent phrase, as has often been observed; an example is (28). In some cases, the implicit argument can control PRO, as in (29).

(28) *Helge lar seg ikke stoppe av hindringer.*
Helge lets REFL not stop by obstacles
'Helge cannot be stopped by obstacles.'

(29) *Pengene lar seg innvinne uten å gå til oppsigelser.*
money-DEF let REFL reclaim without to go to layoffs
'The money can be reclaimed without going to layoffs.'

However, the parallel to regular passives is less than perfect, because LA does not have an external argument, and the implicit agent can only be associated with the second verb.

4 The role of the complex predicate

We have seen that reflexive LA sentences share important properties with regular passives (like the corresponding German construction, see e.g. Pitteroff (2014)). This would be difficult to account for if we simply say that they are active.

Before discussing the question of voice in reflexive LA sentences further, it is necessary to establish another aspect of their analysis. There seems to be consensus that reflexive LA sentences (like the prisoner sentences) are complex predicate constructions (see Taraldsen (1983), Taraldsen (1991), Vikner (1987) and Pitteroff (2014) on Norwegian, Danish and German). The two verbs in reflexive LA sentences behave as one predicate together. This predicate takes one single set of syntactic functions, and behaves as one unit for grammatical rules that operate on argument structure, such as the presentational focus rule (see sentence (12) above).

LA in reflexive sentences is a light verb. We have seen that it has no external argument. The only position in its argument structure is an open position for the argument structure of the second verb. This means that the second verb contributes all the thematic roles that are realized as syntactic functions. When this open position is filled in, we have the argument structure of the complex predicate as a whole. For *la seg stoppe* 'let REFL stop' in a sentence such as (28), the representation will be as in (30) in Lexical Mapping Theory.

(30) la seg < stoppe < agent theme > >
-o -r
OBL SUBJ

Passives of complex predicates — called "long passives" — raise challenges for our understanding of the passive. Examples of long passives are the German (31) (from Wurmbrand 2001, p. 19), and the Norwegian (32).

(31) *dass der Traktor zu reparieren versucht wurde*
that the-NOM tractor to repair tried was
'that they tried to repair the tractor'

(32) *Har mye som må huskes å gjøre.*
have much that must remember-PASS to do
'(I) have many things that I must remember to do.'

Long passives such as (31) and (32) introduce a mismatch between syntax and morphology. A complex predicate can passivize as a whole, but passive morphology is realized on the first verb only. This situation creates a potential problem for the requirement that there must be morphological marking of the passive. Consider sentence (32) above. The first verb *huskes* 'remember-PASS' is uncontroversially passive, but what about the second verb *gjøre* 'do'? It seems to be difficult to say that this verb is active. Its internal argument is realized as a subject, and its external argument is not realized. What is special is of course that its external argument is identified with the external argument of the first verb in the formation of the complex predicate; this is indicated by the indices on the agents in (33).

(33) huske å gjøre 'remember to do' < agent$_i$ < agent$_i$ patient > >

The verb in question is the second part of a complex predicate. The complex predicate is passivized as a whole, and there is only one passivization involved. This passivization is morphologically realized on the first verb only in (31) and (32).

Long passives have the same options of subject choice as other passives (Lødrup 2014). For example, the pseudopassive is possible, as in (34). The choice of the second verb involves the same exceptions as morphologically passive verbs; an example is (35).

(34) *En slik situasjon bør forsøkes å gjøre noe med.*
a such situation ought.to try-PASS to do something with
'One should try to do something about this kind of situation.'

(35) **Elvis bør ikke forsøkes å ligne.*
Elvis should not try-PASS to resemble
'One should not try to resemble Elvis.' [intended]

In Norwegian and some other languages, the second verb of a long passive can have passive morphology, as in (36) (see Lødrup 2014; Haff and Lødrup 2016; Wurmbrand and Shimamura 2017). However, this does not affect the argument. Passive morphology on the second verb has been seen as a kind of verbal feature agreement, licensed by feature sharing in functional structure (see Niño 1997; Sells 2004; Lødrup 2014).

(36) *Dette må forsøkes å gjøres.*
this must try-PASS to do-PASS
'We must try to do this.'

It seems to be difficult to avoid the conclusion that the second verb in a long passive must be considered a passive verb, independently of its own morphology. Its voice is expressed unambiguously, if indirectly, on the first verb in the long passive construction.

5 The status of *la seg*

We concluded with the grammatical tradition that reflexive LA sentences must be passive in some way. The question is then how this should be implemented. This task is in one sense too difficult — these sentences have been discussed a number of times, and there seems to be no simple solution. In another sense, the task is too easy. The reflexive LA construction is special by any account. In Norwegian, it has no clear synchronic relation to other uses of the verb LA, or to other verbs. This means that any account of its properties has to involve at least some idiosyncratic information.

Reflexive LA shares an important property with regular passive verbs: it has no external argument that requires realization as a subject. A difference is that the implicit agent of the complex predicate comes from the second verb, as mentioned in Section 3.

If reflexive LA sentences are passive, the question is what it is that is passive about them. The literature has focused on the second verb. For example, Pitteroff (2014, p. 107) and (2015, p. 45) assume that it is the embedded VP that is passive. A premise of his analysis is a Minimalist conception of complex predicates in which the first verb selects a VP which is 'small' in the sense that it lacks functional projections (Wurmbrand 2001; Cinque 2006).

Within an LFG conception of complex predicates, it would not be natural to assume that two verbs that differ in voice could constitute a complex predicate in a monoclausal structure (pace Lødrup 1996). The reason is that complex predicates behave as units with respect to rules that operate on argument structure. It would be more natural to assume that the whole construction is one passive complex predicate.

This assumption has the consequence that we have to think of reflexive LA as a passive verb, whose passive voice also scopes over the second verb in a complex predicate

construction. This might seem an unintuitive and contrived idea. Is there again a passive verb without passive marking? If there is, the problem now concerns one single verb. Besides, one might consider the empty reflexive a grammatical marker for the passive. It is well known from various languages that simple reflexives can be used to mark different kinds of valency reduction, not only anti-causatives, e.g. Norwegian (37), and middles, e.g. German (38), but also passives, e.g. French (39).[2]

(37) *En dør åpner seg.*
a door opens REFL
'A door opens.'

(38) *Das Buch liest sich leicht.*
the book reads REFL easily
'The book reads easily.'

(39) *Tout se vend ici.*
everything REFL sells here
'Everything is sold here.'

In Norwegian, the reflexive is not used to mark the passive in other cases. If one assumes that it exceptionally functions as a passive marker here, reflexive LA is a passive verb that has no direct counterpart in the active (like English *rumored*, which only exists in the passive).

6 The form of the second verb

The point of departure for this article was the lack of passive morphology with the second verb in reflexive LA sentences. The question is now if passive morphology can be used at all with the second verb. This question raises some problems of analysis. Consider a sentence such as (40). This sentence has a human subject, and can be understood as causative. As a causative, it is a regular subject-to-object-raising sentence. The raised object is then accidentally reflexive, and the sentence is not relevant in this context.

(40) *Bee lar seg bli polstret under.*
Bee lets REFL become padded underneath
'Bee lets herself be padded underneath.' [i.e. her pants are padded]

Sentences such as (41)–(42) are different. These sentences have inanimate subjects, and cannot be interpreted as causative or permissive. I assume that these sentences can only have the same structure as the reflexive LA sentences discussed above.

2 In the literature on German, *sich lassen* 'REFL let' has been seen as the anti-causative of *lassen* 'let' as used in prisoner sentences (e.g. Pitteroff 2014). The reflexive is then the anti-causative marker. Even if this is an intuitive idea, it would be difficult to make use of in a synchronic account of Norwegian.

(41) *Dette lar seg ikke gjøres lenger.*
this lets REFL not do-PASS anymore
'This cannot be done anymore.'

(42) *En del antibiotika lar seg også produseres syntetisk.*
a part antibiotics lets REFL also produce-PASS synthetically
'Some antibiotics can be produced synthetically.'

We see, then, that the second verb in reflexive LA sentences can have active or passive morphology. This is the same phenomenon that was shown in examples (32), (34) and (36) above: The second verb of a long passive can have active or passive morphology in Norwegian. Some reflexive LA sentences might be a bit marginal with a passive second verb in Norwegian, but examples can be found in texts. It is interesting that a passive second verb is normal in Swedish. This difference between the languages was pointed out in Hulthén (1944, p. 199-201). Klingvall (2012) gives Swedish examples such as (43).

(43) *Kakan låter sig bakas med lätthet.*
cake-DEF lets REFL bake-PASS with ease
'The cake bakes easily.'

With the analysis given here, the passive form of the second verb in sentences such as (41), (42) and (43) must be seen as a kind of verbal feature agreement, like in (36) above. It is a general phenomenon that complex predicate constructions in Norwegian can show agreement for certain verbal features — with some variation between speakers. Aagaard (2016, p. 84) shows that participle agreement is possible in reflexive LA sentences.[3] A www example is (44).

(44) *Det hadde ikke latt seg gjort om opptaket var i jpeg.*
it had not let REFL done if recording-DEF were in jpeg
'It would have been impossible if the recording were in jpeg format.'

Examples (36) and (44) might look like cases of suffix copying. However, it is important that verbal feature agreement is a more "abstract" phenomenon. Voice agreement concerns the grammatical feature passive, and does not require the two verbs to mark the passive in the same way (Lødrup 2014). This is shown in (45) and (46), in which

3 Imperative agreement is not uncommon in complex predicates, as in *Forsøk å stupe/stup!* 'try-IMP to dive-INF/dive-IMP' (Havnelid 2015). It is striking, then, that imperative agreement is completely unacceptable in reflexive LA sentences: *La deg ikke stoppe/*stopp!* 'let-IMP REFL not stop-INF/stop-IMP'. This must be related to the fact that the second verb of a reflexive LA sentence is not a real active.

complex predicates with voice agreement each have one morphological passive and one periphrastic passive.[4]

(45) *Deponiet foreslås å bli lagt til et område som …*
depot-DEF suggest-PASS to become placed to an area that
'They suggest that the depot be placed in an area that … '

(46) *Viktige stridsspørsmål blir unnlatt å presiseres.*
important issues become neglected to clarify-PASS
'They neglect clarifying important issues.'

7 Conclusion

A passive verb must have an identifiable expression of its passive voice. However, this expression does not necessarily have to be on the verb itself. For the second verb in a complex predicate it is enough that the first verb is marked.

The reflexive LA construction is special on all accounts. It has no clear synchronic relation to similar constructions, or to other uses of the verb LA in Norwegian. I have argued that its properties are best accounted for when we assume that *la seg* and its second verb constitute a passive complex predicate.

8 Acknowledgements

I was lucky to have Helge as a colleague when I worked in Bergen in the 1980s. I am grateful for what I have learned from him, both then and later. A highlight of my early linguist days was his dissertation defense. (It is remarkable how his dissertation anticipated aspects of LFG and HPSG.) I never stop laughing at his classical poems. We have also had some laughs together.

References

Aagaard, Teodor Ekblad (2016). *Doble partisipper i norsk: Verbal trekkongruens og restrukturering*. MA thesis. University of Oslo.

Åfarli, Tor A. (1992). *The Syntax of Norwegian Passive Constructions*. Amsterdam: John Benjamins.

Åfarli, Tor A. and Kristin Melum Eide (2003). *Norsk generativ syntaks*. Oslo: Novus.

4 A possible objection is that voice agreement in long passives is very common in Norwegian. Why is it then unusual in reflexive LA sentences? It must be considered that different complex predicates can behave in different ways with respect to voice agreement. For example, French has long passives of aspectual verbs (such as *finir* 'finish') with or without voice agreement, while long passives of other verbs (such as e.g. *tenter* 'try', *oublier* 'forget') seem to require voice agreement (Haff and Lødrup (2016), Chantal Lyche pc).

Baker, Mark (1993). "Why unaccusative verbs cannot dative-shift". In: *Proceedings of the North East Linguistic Society 23*. Ed. by Amy J. Schafer. Department of Linguistics, University of Massachusetts, pp. 33–47.

Cinque, Guglielmo (2006). *Restructuring and Functional Heads: The Cartography of Syntactic Structures*. Oxford: Oxford University Press.

Comrie, Bernard (1976). "The syntax of causative constructions: Cross-language similarities and divergences". In: *Syntax and Semantics 6: The Grammar of Causative Constructions*. Ed. by Masayoshi Shibatani. New York: Academic Press, pp. 261–312.

Drummond, Alex and Dave Kush (2015). ""Reanalysis" is raising to object". In: *Syntax* 18.4, pp. 425–463.

Dyvik, Helge (1980). "Har gammelnorsk passiv". In: *The Nordic Languages and Modern Linguistics 4*. Ed. by Even Hovdhaugen. Oslo: Universitetsforlaget, pp. 82–107.

Engdahl, Elisabet (2006). "Semantic and syntactic patterns in Swedish passives". In: *Demoting the Agent. Passive, Middle and Other Voice Phenomena*. Ed. by Benjamin Lyngfelt and Torgrim Solstad. John Benjamins, pp. 21–45.

Faarlund, Jan Terje, Svein Lie, and Kjell Ivar Vannebo (1997). *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Falk, Hjalmar and Alf Torp (1900). *Dansk-norskens syntax i historisk fremstilling*. Aschehoug.

Gunkel, Lutz (1999). "Causatives in German". In: *Theoretical Linguistics* 25.2/3, pp. 133–160.

Haff, Marianne Hobæk and Helge Lødrup (2016). "Où en est le «passif long» en français?" In: *Syntaxe et sémantique* 1, pp. 153–173.

Haspelmath, Martin (1990). "The grammaticization of passive morphology". In: *Studies in Language* 14.1, pp. 25–72.

Havnelid, Inga Margrethe (2015). *"Prøv å fakk skurken som gjemmer seg på Manhattan": Om dobbel imperativ i norsk*. MA thesis. University of Oslo.

Herslund, Michael (1986). "Causatives, double objects, and the ergativity hypothesis". In: *Scandinavian Syntax: Papers from the Ninth Scandinavian Conference of Linguistics*. Ed. by Östen Dahl. Centre for Languages and Literature, Lund University, pp. 142–153.

Hulthén, Lage (1944). *Studier i jämförande nunordisk syntax*. Göteborgs högskolas årsskrift 50.

Keenan, Edward L. and Matthew S. Dryer (2007). "Passive in the world's languages". In: *Language Typology and Syntactic Description 1. Clause Structure*. Ed. by Timothy Shopen. Cambridge University Press, pp. 325–361.

Keyser, Samuel Jay and Thomas Roeper (1984). "On the middle and ergative constructions in English". In: *Linguistic Inquiry* 15.3, pp. 381–416.

Klingvall, Eva (2012). "How to approach complex passives". In: *Discourse & Grammar. A Festschrift in Honor of Valéria Molnár*. Ed. by Johan Brandtler, David Håkansson,

Stefan Huber, and Eva Klingvall. Centre for Languages and Literature, Lund University, pp. 395–410.

Labelle, Marie (2013). "Anticausativizing a causative verb: The passive *se faire* construction in French". In: *Non-Canonical Passives*. Ed. by Artemis Alexiadou and Florian Schäfer. Amsterdam: John Benjamins, pp. 235–260.

Levin, Beth and Malka Rappaport Hovav (1995). *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Cambridge, MA: MIT press.

Lødrup, Helge (1996). "The theory of complex predicates and the Norwegian verb få 'get'". In: *Working Papers in Scandinavian Syntax* 57, pp. 76–91.

– (2000). "Exceptions to the Norwegian passive: Unaccusativity, aspect and thematic roles". In: *Norsk lingvistisk tidsskrift* 18.1, pp. 37–54.

– (2014). "Long passives in Norwegian: Evidence for complex predicates". In: *Nordic Journal of Linguistics* 37.3, pp. 367–391.

Niño, María-Eugenia (1997). "The multiple expression of inflectional information and grammatical architecture". In: *Empirical Issues in Formal Syntax and Semantics*. Ed. by Francis Corblin, Danièle Godard, and Jean-Marie Marandin. Peter Lang, pp. 127–47.

Pitteroff, Marcel (2014). "Non-canonical *lassen*-middles". PhD thesis. University of Stuttgart.

– (2015). "Non-canonical middles: a study of personal *let*-middles in German". In: *Journal of Comparative Germanic Linguistics* 18.1, pp. 1–64.

Platzack, Christer (1986). "The structure of infinitive clauses in Danish and Swedish". In: *Scandinavian Syntax. Workshop at the Ninth Scandinavian Conference of Linguistics*. Ed. by Östen Dahl and Anders Holmberg. Institute of Linguistics, University of Stockholm, pp. 123–137.

Reis, Marga (1973). "Is there a rule of subject-to-object raising in German?" In: *Papers from the 9th Regional Meeting of the Chicago Linguistic Society*. Ed. by Claudia Corum, T. Cedric Smith-Stark, and Ann Weiser. Chicago Linguistic Society. Chicago, pp. 519–529.

Sells, Peter (2004). "Syntactic information and its morphological expression". In: *Projecting Morphology*. Ed. by Louisa Sadler and Andrew Spencer. CSLI Publications, pp. 1287–225.

Siewierska, Anna (1984). *The Passive: A Contrastive Linguistic Analysis*. London: Croom Helm.

Taraldsen, Knut Tarald (1983). *Parametric Variation in Phrase Structure: A Case Study*. Dissertation, University of Tromsø.

– (1991). "A directionality parameter for subject-object linking". In: *Principles and Parameters in Comparative Grammar*. Ed. by Robert Freidin. Cambridge, MA: MIT Press, pp. 219–268.

Vikner, Sten (1987). "Case assignment differences between Danish and Swedish". In: *Proceedings of the Seventh Biennial Conference of Teachers of Scandinavian Studies in Great Britain and Northern Ireland*. Ed. by Robin Allan and Michael P. Barnes. University College London, pp. 262–281.

Wurmbrand, Susanne (2001). *Infinitives: Restructuring and Clause Structure*. Berlin: Mouton de Gruyter.

Wurmbrand, Susanne and Koji Shimamura (2017). "The features of the voice domain: Actives, passives, and restructuring." In: *The Verbal Domain*. Ed. by Roberta d'Alessandro, Irene Franco, and Ángel Gallego. Oxford: Oxford University Press, pp. 179–204.

From LFG structures to dependency relations

Paul Meurer

Abstract. In this article, I describe the derivation of dependency structures from LFG analyses, with a focus on the Norwegian grammar NorGram. Although it is the f-structures that at a first glance resemble dependency structures most, I show that c-structures are the correct starting point for the conversion, and I outline a conversion algorithm that relies on information from both c- and f-structure, the projection operator, and the grammar itself. The derived dependency structures are projective with non-atomic relations, but can be converted into non-projective dependencies with atomic relations, and further into Universal Dependency-style structures. As an application, I describe how derived dependency versions of the NorGramBank gold-standard treebank are used to train dependency parsers with acceptable precision.

1 Introduction

Lexical Functional Grammar (Bresnan 2001) is a theoretically motivated grammar formalism that allows the encoding of a very rich set of grammatical information. This is exemplified by the Norwegian LFG grammar NorGram[1], which has been used to build a large treebank of automatically parsed and disambiguated sentences (NorGramBank), including a smaller gold-standard treebank of manually disambiguated analyses (Dyvik et al. 2016).

In contrast, many existing larger treebanks are manually or semi-automatically constructed,[2] and they are expressed in more light-weight and less theory-driven formalisms, such as phrase-structure trees of the Tiger treebank type, or Dependency Grammar. The latter formalism has recently gained much attention and popularity, most notably through the Universal Dependencies (UD) initiative.[3] The UD project seeks to provide dependency treebanks for many languages (currently comprising 64 treebanks for 47 languages) in a comparable way, by using 'universally' agreed-on and accepted coding guidelines and tagsets, while at the same time trying to keep a sensible balance between divergent design goals (De Marneffe et al. 2014; Nivre, Marneffe, et

1 https://clarino.uib.no/redmine/projects/inesspublic/wiki/NorGram_documentation

2 Notable exceptions are the Redwoods and similar HPSG treebanks, which are constructed in a way similar to NorGramBank, and the Alpino dependency treebank.

3 http://universaldependencies.org

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1341.

al. 2016). Among those goals are ease of comprehension also for non-linguists such as language learners and engineers, on the one hand, and suitability for computer parsing with high accuracy, on the other hand. The layout and distribution format of the UD treebanks is such that they can readily be fed into a training pipeline for a statistical parser, e.g. MaltParser (Nivre, Hall, et al. 2006), MATE (Bohnet 2010) or the Stanford Neural Network parser (Chen and Manning 2014).

Even though it is still an open question how well such derived statistical parsers perform compared to hand-crafted grammars, the idea is compelling: training a statistical parser from an existing treebank is much less time-consuming than developing a broad-coverage rule-based grammar. In addition, statistical parsers tend to be more robust and operate at a much higher speed than rule-based (e.g. LFG or HPSG) grammars.[4] Even though statistical parsers cannot compete with detailed hand-crafted computational grammars in terms of depth of linguistic analysis and richness of detail, they are nevertheless potentially more suited for certain classes of applications where a fine-grained syntactic analysis is not necessary, and speed and coverage are of higher importance. Among such applications are data mining and information extraction of various kinds.

Motivated by such considerations, and the desire to create a consistently annotated, UD-compatible dependency treebank with relatively little effort, I describe in this article a conversion algorithm from LFG to dependency structures of various types, and I present the resulting dependency treebank and a spin-off product, a set of dependency parsers for Norwegian Bokmål.

Two quite similar approaches to the conversion of LFG structures into dependency structures have been described in (Øvrelid, Kuhn, et al. 2009) and (Çetinoğlu et al. 2010). Below, I will compare their approaches to the one I have chosen. The admittedly more complex task of converting a dependency treebank into a treebank of LFG structures has also been performed (Haug 2012). Crucial to the success was the availability of structural relations in the dependency structures of that particular treebank, the PROIEL treebank (Haug and Jøhndal 2012), that go beyond what is generally coded in dependency structures, namely secondary edges.

When going from LFG to dependency structures, enough structure should be available to allow the construction of the needed dependency relations. The question is mainly which part of the rich LFG structure (c-structure, f-structure, the projection relation between them) to base the conversion on, and which information to discard.

At a first glance, it is the f-structures that resemble dependency structures most. Dependency structures can roughly be seen as impoverished f-structures, where all attributes except the functional ones, corresponding to dependency relations, and all structure sharing have been removed. Both Øvrelid, Kuhn, et al. (2009) and Çetinoğlu

4 This does not apply to dependency parsers that operate in the (rule-based) Constraint Grammar framework (Karlsson 1990).

et al. (2010) use this correspondence as the starting point for their conversions. This correspondence is however not perfect; f-structure PRED values cannot easily be related to surface words (which the dependency nodes should consist of), because the projection relation is not injective.

Øvrelid, Kuhn, et al. (2009) solve this problem by introducing generic co-head edges between the surface form contributing the PRED value of the projected f-structure and other surface forms that project to the same f-structure. Çetinoğlu et al. (2010) describe a similar approach: they construct a modified f-structure, where every surface node corresponds to a proper PRED value. However, they give no detailed account of their algorithms.

In contrast, I chose a conversion that starts with the c-structure, but exploits the f-structure and the projection operator to arrive at the correct dependency relations and labels.

The dependency relations that are the result of the algorithm that will be outlined in Section 2 are peculiar in that they inherit a characteristic of the c-structures they are derived from: they are projective, which c-structures trivially are. This entails that the derived dependency relation labels are non-atomic in general; they are the concatenation of basic grammatical function relations, e.g. in the case of long-distance dependencies. An additional transformation has to be performed to make all relations basic, at the expense of projectivity, in order to arrive at traditional dependency structures. Both representations are equivalent and can be transformed into each other.

Even these non-projective dependency structures are quite different from dependency structures that adhere to the Universal Dependencies coding guidelines. Our derived dependency structures basically inherit their head-dependent relationship from the functional relations in the LFG f-structure, which for NorGram analyses entails that function words like (non-selected) prepositions, auxiliaries,[5] modals and coordinations (but not complementizers) are heads, having content words as dependents. These structures resemble quite closely the dependencies of the PROIEL–TOROT–Menotec family of treebanks[6] and the German TüBa-D/Z treebank,[7] among others (disregarding relation names), although TüBa-D/Z treats coordination differently and more in line with UD. In the UD initiative, on the other hand, a guiding principle is that heads should be content words, whereas function words modify the head words. This design decision was made to achieve a high degree of parallelism between dependency structures of different languages.

In Section 2.6, I will outline how the non-projective dependency structures derived from LFG analyses can be converted into UD-compatible structures.

5 Other LFG grammars might treat auxiliaries differently; e.g., in the English Pargram grammar, they have no PRED value on their own and merely contribute a feature to the f-structure.

6 See http://clarino.uib.no/iness

7 http://www.sfs.uni-tuebingen.de/de/ascl/ressourcen/corpora/tueba-dz.html

Applying the mentioned conversion algorithms to an existing gold-standard LFG treebank for Norwegian Bokmål, I derived a set of three dependency treebanks: one consisting of projective structures, a second having non-projective structures with atomic labels, and a third, UD-conformant dependency treebank.

All three treebanks were used to train statistical dependency parsers using the Stanford Neural Network parser framework and the MATE parser tools.

The performance of the resulting parsers is comparable to the numbers mentioned in the literature, with some interesting differences. I have, however, not tried to fine-tune the tagset and the training parameters in order to maximize performance. Training could also benefit from an improved gold-standard corpus.

In the conclusion, I briefly mention an application the trained parser has already found in the domain of quotation extraction.

2 From c-structure to dependency

As stated in the introduction, although f-structures conceptually resemble dependency structures (by interpreting sub-f-structures as dependency nodes and their PRED values as node labels, and taking f-structure attributes as dependency relations), they cannot be converted into traditional bilexical dependency structures without resorting to c-structure and projection information. There are several reasons for this.

Firstly, linear order information is not coded in f-structures, whereas the ordered nodes of a dependency tree should mirror the surface word order of the analyzed sentence. An ordering on sub-f-structures can only be imposed by relating them to nodes in the ordered c-structure via the projection operator, and it is in many cases far from obvious how this should be done. In addition, dependency node labels are exactly the surface token strings of the sentence. F-structure PRED values, on the other hand, are in most cases the dictionary entry forms of inflected surface words, or other abstractions from the surface form. Here, the correct surface word form might be recoverable by making use of the projection operator and other information coded in the internal representation of the LFG analysis. In some cases, however, there is not even an easily discernable trace of a surface form in the f-structure at all. These problems are exemplified in Figure 1, which displays a tentative dependency structure derived from an f-structure for sentence (1). The possessive *min* 'my' projects to the predicate *pro*, the selected preposition *om* 'about' is fused with the verb predicate *drømme*om* 'dream about', and the demonstrative *denne* 'this' even gives rise to two PRED values.

(1) *Min katt drømmer om dette.*
my cat dreams about this
'My cat dreams about this.'

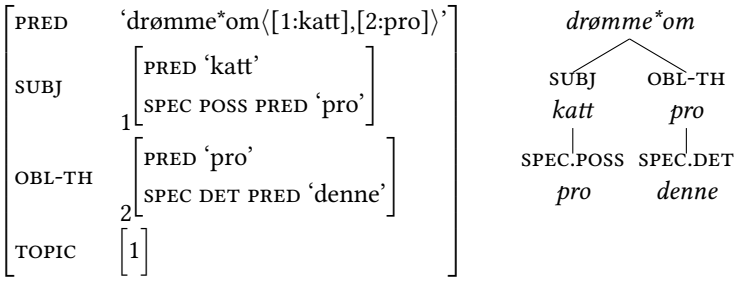


Figure 1: F-structure and derived dependency graph for (1)

For all these reasons, a more promising approach could be to derive dependency relations directly from c-structures, using f-structure information solely to construct relation labels.

2.1 The basic Lifting algorithm

In order to be able to state the algorithm that accomplishes this derivation (the Lifting algorithm), we need to recall the notion of functional head (in the c-structure!). A daughter node Y of a node X is a *functional head* (also called *f-structure head*) if Y is annotated with the equation $\uparrow{=}\downarrow$, or equivalently, Y and X share their features, or Y and X project to the same f-structure. This notion is different from the X'-theory concept of a c-structure head (e.g., N is the c-structure head of NP).

Let us first assume that every non-terminal c-structure node X has exactly one functional head Y. Under this assumption, the Lifting algorithm is easy to formulate:

Lifting algorithm, basic version.

1. Recursively replace each non-terminal node by its functional head node. In other words, lift each functional head node up to its mother node (which it replaces). Since we assume that each non-terminal node has exactly one functional head, this procedure is well-defined.

2. Label the edge between node X and daughter node Y with the f-structure path from $\varphi(X)$ to $\varphi(Y)$, where $\varphi : C \rightarrow F$ is the projection operator. If there is more than one path (because of structure sharing in the f-structure) choose the path that consists of grammatical functions (that is, contains no discourse functions like TOPIC *or* FOCUS*, but rather* SUBJ, OBJ, ADJUNCT *etc.*[8]*). If there is more than one such path, choose the shortest one. This is called the* minimal path.

8 The complete list of grammatical functions in NorGram is: SUBJ, OBJ, OBJ-TH, OBJ-BEN, OBL, OBL-TH, OBL-COMPAR, PREDLINK, COMP, XCOMP, X, ADJUNCT, NULL.

Figure 2 shows some steps in the application of the basic Lifting algorithm for example (2).[9]

(2) *Hunden sover.*
the dog sleeps
'The dog sleeps.'

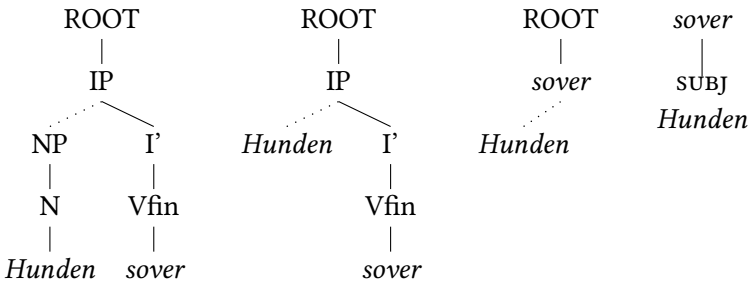


Figure 2: Steps in the application of the basic Lifting algorithm for the sentence (2)

2.2 Finding heads

In fact, all but the most basic c-structures violate the assumption of the basic lifting algorithm: a c-structure can contain nodes with multiple functional heads or without functional heads. Therefore, the Lifting algorithm has to be extended to such c-structures.

In case of multiple functional heads, the idea is to turn all but one of them into dependents, according to configurational details in the projected f-structure. We first take a closer look at c-structure co-heads.

Most c-structure terminal nodes (word forms) straightforwardly project to an f-structure whose PRED value is the associated semantic form (the base form of nouns and adjectives, and the infinitive with its subcategorization frame in the case of verbs). This is expressed via an LFG functional equation of type (3) associated to the word form, here *hund* 'dog'.

(3) (↑ PRED) ='hund'

In some cases, however, the PRED value is embedded deeper in the f-structure the surface node projects to. This is true for determiners, whose PRED value is embedded in the projected f-structure along the path SPEC DET, via equation (4). The same holds for quantifiers, which are embedded along SPEC QUANT, and possessives (SPEC POSS).

(4) (↑ SPEC DET PRED) = 'denne'

9 In this and subsequent figures, straight lines are drawn between nodes and their functional heads, whereas other c-structure edges are dotted lines.

Determiners (as well as quantifiers and possessives) are also c-structure functional heads, and their c-structure complements are f-structure co-heads, such that a c-structure fragment (5) corresponding to the phrase *denne hunden* 'this dog', where all nodes lie in the same functional domain, projects the f-structure (6). (Only the relevant parts are shown.)

(5) DP → D NP

(6)
$$\begin{bmatrix} \textsc{pred} & \text{'hund'} \\ \textsc{spec} & \begin{bmatrix} \textsc{det} & \begin{bmatrix} \textsc{pred} & \text{'denne'} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The path from the projected f-structure to the associated PRED value I call the *embedding path* (SPEC.DET in the above example). The embedding path is empty when the PRED value is at the top level of the projected f-structure. Clearly, among several co-head nodes, we wish to turn those nodes that have a non-trivial embedding path into dependents of the node with empty embedding path (if it exists), and their relations will basically be their embedding paths. Accordingly, in the example above, the constructed relation will be (7).

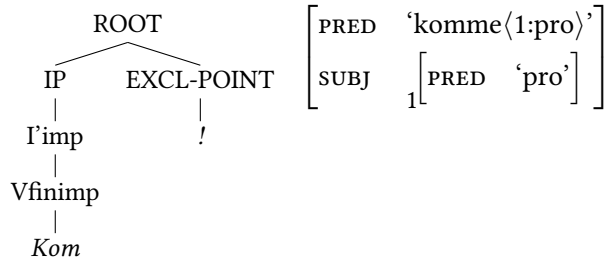
(7)

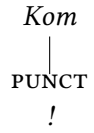

As we have seen, the embedding path of a lexical node cannot be deduced from the c- and f-structures and the projection operator alone. Rather, it has to be extracted from the functional equations attached to the lexical entry in the LFG grammar. Therefore, the details of the outlined algorithm are dependent on the grammar the sentence was parsed with.

There may also be lexical elements in the c-structure that are functional co-heads, but have no corresponding PRED value in the f-structure at all. Examples are selected prepositions and punctuation marks. Here, we arbitrarily construct an embedding path, which in the punctuation case will simply be PUNCT, as in (8, 9) for the sentence *Kom!* 'Come!'.[10] Selected prepositions will be dealt with below.

Even though coordinations are functional heads not associated with a PRED value, there is no need to give them special treatment, since they are never co-heads.

10 This example illustrates how the explicit subject information from the f-structure is lost in the conversion.

(8)


(9)


Now we are in the position to formulate the extended lifting algorithm:

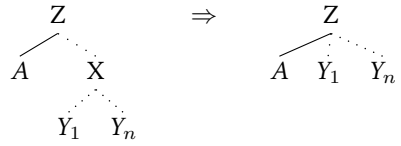
Lifting algorithm, extended version.

1a. If none of the daughter nodes $Y_1, \ldots Y_n$ of node X is a functional head, replace X by the daughter nodes as direct children of the parent Z of X, as illustrated in (10). Then proceed as before.

1b. If more than one daughter node of X is a functional head, select the node with shortest or empty embedding path as replacement for X. The remaining nodes will be treated as dependents, their relations to X being their embedding paths.

1c. As a last resort, if there is more than one such node, select the first of them as replacement.

2. Label the edge between node X and daughter node Y with the minimal f-structure path from $\varphi(X)$ to $\varphi(Y)$, concatenated with the embedding path of Y.

(10)


It is immediate from the construction that the dependency relations constructed in this way are *projective* dependencies; the sequence of words reachable from a given node along dependency arrows has no gaps, and there are no crossing edges.

Figure 3 shows the application of both *1a.* and *1b.* for sentence (11).

(11) *I dag sov noen barn lenge.*
today slept some children long
'Today some children slept long.'

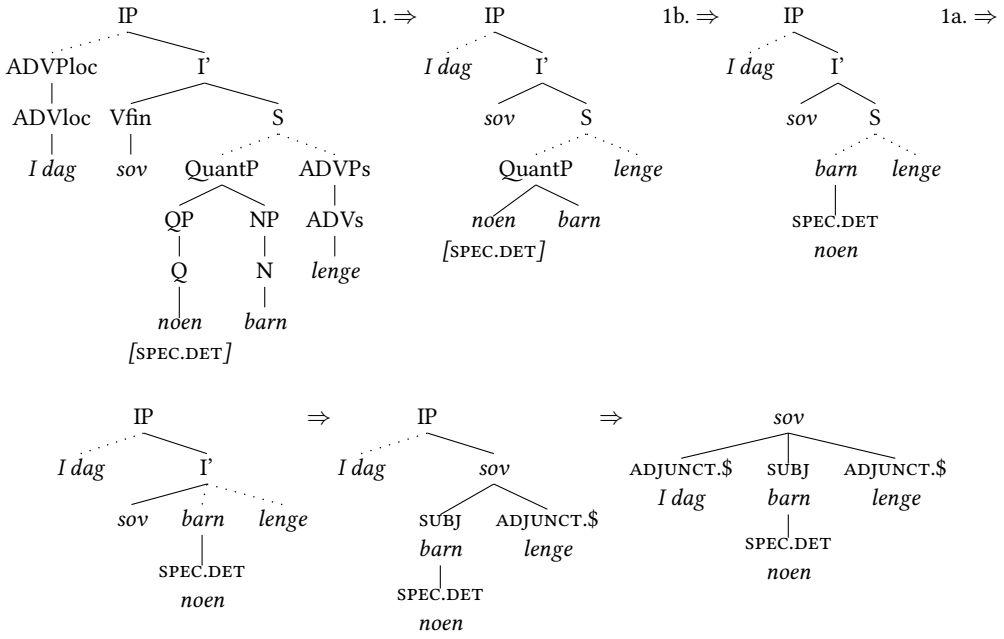


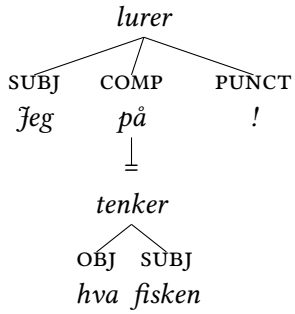
Figure 3: Steps in the application of the extended Lifting algorithm for the sentence (11). Nontrivial embedding paths are shown in brackets.

2.3 Words without PRED values

Selected prepositions have no semantic value on their own and hence do not contribute a PRED to the f-structure, and they are functional co-heads. Such a selected preposition gets an empty embedding path, which makes it the dependency head of the subordinate clause, according to rule *1c*. The relation is labeled '=', indicating that the selected preposition is not connected with a grammatical function, but merely acts as a mediator between e.g. the COMP relation and the subordinate clause, as in the analysis (13) of sentence (12).

(12) *Jeg lurer på hva fisken tenker.*
I wonder about what the-fish thinks
'I am wondering what the fish is thinking.'

(13)



Commas are treated in the same way, for similar reasons.

Complementizers including the infinitival marker *å* have no PRED value either; they merely contribute a COMP-FORM value to the f-structure. Here, the corresponding c-structure labels (Cnom, Cinf etc.) are chosen as embedding paths.

2.4 Projective vs. non-projective dependencies

As we have seen, the dependency structures derived by the Lifting algorithm are projective dependencies that potentially have compound relation labels $R = R_1.R_2 \ldots R_n.S$, consisting of the concatenation of more than one grammatical function or set inclusion R_i, and a possibly empty suffix S deriving from lexical embeddings.

This is not what dependency structures should look like. It is, however, relatively straightforward to transform such projective dependencies into non-projective dependencies without compound relation labels. The main idea is to move along the component relations of a compound relation to find the new head of a dependent. Given a relation $X \xrightarrow{R_0.R_1 \ldots R_n.S} Y$, there must also exist an atomic relation $X \xrightarrow{R_0} X_1$. The node X_1 corresponds to the c-structure surface node that gives rise to the PRED value in the sub-f-structure along R_0 of the projection of X. Then we can replace the relation $X \xrightarrow{R_0.R_1 \ldots R_n.S} Y$ by the relation $X_1 \xrightarrow{R_1 \ldots R_n.S} Y$, and by applying this process recursively we end up with a dependency structure $X_n \xrightarrow{R_n.S} Y$ (or $X_{n+1} \xrightarrow{S} Y$) without compound relation labels.

In the projective dependency structure for sentence (14) in Figure 4, there are two compound relations.[11]

(14) *Dette vet jeg ikke hva jeg skal si til.*
this know I not what I shall say to
'This I don't know what to say about.'

The second of them, '*hva* $\xleftarrow{\text{XCOMP.OBJ}}$ *skal*', is resolved by replacing it with a relation starting from the target '*si*' of the XCOMP relation, namely '*hva* $\xleftarrow{\text{OBJ}}$ *si*'. The

11 In this and the following examples, linear display mode is chosen for the dependency analyses. This mode is more suitable for longer examples, and it makes non-projectivity immediately visible through crossing edges.

other compound relation is resolved recursively, finally resulting in the relation '*Dette* $\xleftarrow{\text{OBJ:SPEC.DET}}$ *til*'.

The outlined procedure assumes that the relation $X \xrightarrow{R_0} X_1$ is unique. However, this can only be guaranteed for functional relations, in virtue of the LFG uniqueness condition. If R_0 denotes set inclusion ('$'), any of the possible relations $X \xrightarrow{\$} X$' (where X' stands for the various set members) could be taken as replacement for $X \xrightarrow{R_0.R_1...R_n.S} Y$. To avoid this problem, we label the set inclusions in the f-structure with unique subscripts.

Finally, the relations are simplified if possible; the set inclusion marker '$' is removed where it can be understood as implicit in the relation (e.g., ADJUNCT.$ is reduced to ADJUNCT since adjuncts are always set-valued), suffixes are dropped (e.g., OBJ:SPEC.DET is reduced to OBJ), and relations from lexical embeddings are simplified (e.g., SPEC.DET is reduced to DET).

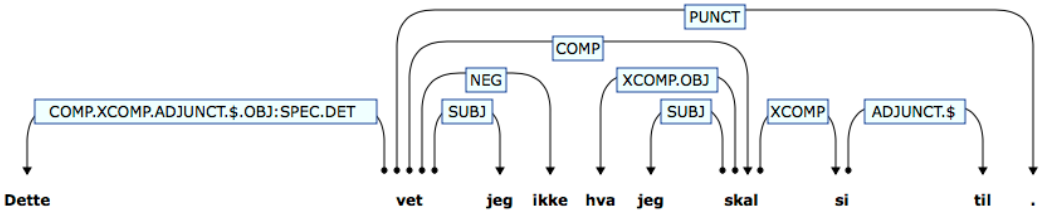


Figure 4: Projective dependency structure with compound labels for (14)

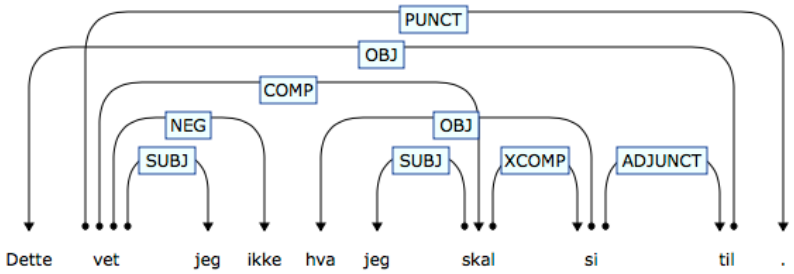


Figure 5: Non-projective dependency structure with atomic labels

2.5 Secondary edges

In some variants of Dependency Grammar, it is possible to include secondary edges, leading to structure sharing and dependency structures that no longer are trees, but directed graphs. Such structures bear an even closer resemblance to f-structures, which also are directed graphs. Secondary edges are typically used to code functionally bound arguments of the subordinate verb in raising or equi constructions, or to code the sub-

ject of a participle that modifies its own subject. In the latter case, the resulting graph is even circular. In f-structures there is also structure sharing arising from discourse functions (e.g., TOPIC, FOCUS), which could be modeled with secondary edges.

The construction of secondary edges, though not difficult in principle, is not yet covered by the described or implemented algorithms.

2.6 Conversion to Universal Dependencies

As mentioned, UD-style dependencies differ from the dependencies described in the previous sections (which I will call *LFG-style dependencies* in what follows) in the treatment of function words: whereas copula verbs, auxiliaries, modals, prepositions and coordinations are heads in the LFG-style dependencies, in the same way as they are LFG functional heads in NorGram[12] (with the exception of coordinations), function words are never heads in UD dependencies; they connect via *functional relations*, in the terminology of the UD project, to the content word. This means in concrete terms that in order to transform LFG-style dependency structures into UD dependencies, the function word head-dependent relations have to be inverted, and otherwise adapted.

This is done in an at least conceptually quite straightforward way by recursively turning a function word (copula, auxiliary, modal, non-selecting preposition) into a daughter node of the content word it heads (which amounts to reversing the head-dependent relation arrow) and turning all remaining daughter nodes of the function word into daughter nodes of the content word.

A special case are coordinations, which in the LFG-style dependencies derived from NorGram analyses are binary branching and are heading two content words (or, recursively, a content word and another binary coordination). In contrast, in UD coordinations, it is the first conjunct that is considered the head of the coordinated phrase, all other conjuncts being dependents. The coordinating conjunctions are attached to the word following them. This configuration is in accordance with the UD principle to give no function word head status; this asymmetric structure is, however, less elegant, as it does not reflect the equal semantic status of the conjuncts in the phrase.

In addition to relation edges, the relation labels too have to be adjusted in this transformation. The current version of the UD standard (UD v.2) recognizes 37 types of syntactic relations, whereas there are around 45 relation types in the LFG-style dependencies. In many cases, the UD and the LFG relations code syntactic functions at a different level of detail, and there is no straightforward mapping from LFG-style to UD-style relations. The concrete replacement of an LFG-style relation can depend on morphological features, c-structure parent labels and f-structure attribute values, in addition to the relation label itself. A typical example is the ADJUNCT relation, which translates to *nmod* if the adjunct is a noun phrase, to *amod* if it is an adjective, to *acl:relcl* if the adjunct is a relative clause, and so on.

12 One should keep in mind that other LFG grammars might treat auxiliaries differently; e.g., in the English ParGram grammar, auxiliaries only contribute with a feature to the f-structure.

The outlined derivation procedure is exemplified in Figure 6, which shows the LFG-style dependency structure for sentence (15) and the remarkably different UD-style dependency structure derived thereof.

(15) *Vi skulle ha kjørt med båt, buss eller bil.*
we should have driven with boat, bus or car
'We should have taken boat, bus or car.'

LFG allows certain types of multi-word expressions (MWEs), such as adverbials (*i dag* 'today'), complex prepositions (*ved siden av* 'next to') or named entities (*Møre og Romsdal*), to be atomic surface nodes. In contrast, MWEs are syntactically analyzed in UD; each component word represents a dependency node. UD distinguishes between three types of MWEs: fixed expressions, flat exocentric semi-fixed expressions, and endocentric analyzable compounds. Since MWEs recognized by NorGram reveal no internal structure, I treat them uniformly as fixed expressions, even though named entity MWEs arguably could be viewed as analyzable compounds with internal heads. Hence, MWEs will be annotated by attaching all non-first components to the first component via the relation *fixed*. This is shown for sentence (16) in Figure 7.

(16) *Sogn og Fjordane ligger ved siden av Møre og Romsdal.*
Sogn og Fjordane lies at the side of Møre og Romsdal
'Sogn og Fjordane is situated next to Møre og Romsdal.'

3 Training a dependency parser

The INESS NorGramBank treebank is a set of treebanks analyzed with the Norwegian LFG grammar NorGram, comprising around 5,5 million sentences (75 million words), of which 4.7 million are in the Bokmål variant of Norwegian. 28,500 of the Bokmål sentences have manually disambiguated and controlled analyses, representing the gold standard[13], whereas the analyses of the remaining sentences are disambiguated using a stochastic disambiguation module (Riezler and Vasserman 2004) that was trained on the gold standard.

As outlined in the previous sections, there are three types of dependency structures that can be derived from LFG analyses: projective and non-projective LFG-style dependencies, and UD-style dependencies. These derivations were applied to the NorGramBank gold standard, which resulted in three corresponding dependency treebanks.

In the experiment that I will describe in this section, those three treebanks were used to train statistical parsers, whose performance was then compared, against each

13 The gold standard contains only correct analyses; sentences where the grammar provides only incorrect analyses were not included.

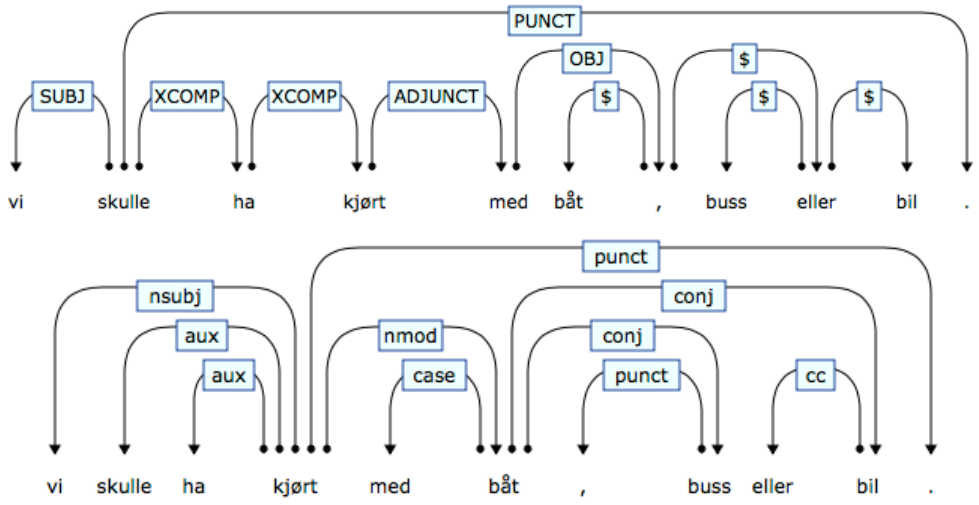


Figure 6: LFG-style and UD-style dependency structure for sentence (15)

other, and against a parser trained on the official Norwegian UD v1.4 treebank (20,000 sentences).[14]

Training of the dependency parsers was done using two different statistical dependency parser frameworks: the Stanford Neural Network Dependency Parser, and the graph-based parser from the MATE tools. As is usual, the treebank sentences were randomly divided into training, development and test sets of relative sizes 8 : 1 : 1 for each treebank. The training set had 18,859 sentences. The same training–test split was used for all three variants of the treebank. The treebanks were exported in variants of the CoNLL format, formats accepted by the parser training Java programs of the two parser frameworks. As POS tagset, the lexical categories of the c-structure nodes were chosen. Since the dependency parsers do not easily accept tokens with whitespace, NorGramBank multi-word expressions had to be split into separate tokens. As POS tags of the component tokens, the lexical category of the multi-word expression was used, extended with an '/MWE' suffix.

The parsers for the Norwegian UD v1.4 treebank were trained and tested on the training, development and test sets included in the release. For both parser frameworks, the standard training settings were used.

In addition to the training and development sets, the training algorithm for the Stanford parser also needs a word embedding file, which contains distributed representations of the words of the language. This word embedding file for Norwegian was

14 See http://universaldependencies.org

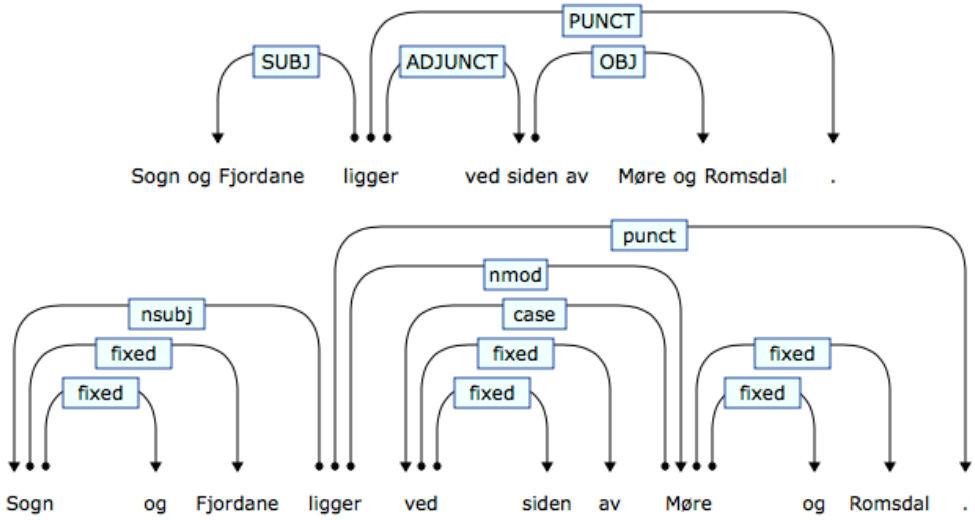


Figure 7: Treatment of multi-word expressions in LFG-style and UD-style dependencies for sentence (16)

created with the program word2vec,[15] using as input the combined word tokens of the corpora *Norwegian Newspaper Corpus* (newspaper text), *forskning.no* (popular science) and *Talk of Norway* (parliamentary debates), with more than 1.58 billion tokens altogether.[16]

The obtained precision for the parsers trained with the different treebanks and parser frameworks is given in Table 1, in terms of *Unlabeled Attachment Score* (UAS) and *Labeled Attachment Score* (LAS). In addition, the percentage of sentences with fully correct (unlabeled and labeled) attachment is given in parentheses. Both training and testing was done on pretokenized and POS-tagged input using the gold-standard POS tagset.

There is a striking difference between the parsers trained with the Stanford framework and with MATE: the MATE parsers perform much better. This is consistent with observations in the literature, where MATE is among the best performing parsers in many comparisons, e.g., Lavelli (2016) and Choi et al. (2015). However, the difference is much higher for the parsers trained on the UD v1.4 Norwegian treebank than for those trained on the LFG-derived treebanks, a fact that I cannot offer an explanation for.

15 https://code.google.com/archive/p/word2vec/

16 See http://clarino.uib.no/corpuscle

	Stanford NN		MATE	
Treebank	UAS	LAS	UAS	LAS
LFG-proj	91.67 (65.72)	89.30 (57.41)	93.81 (71.51)	90.82 (**61.58**)
LFG-nonproj	91.04 (62.44)	89.30 (57.11)	93.63 (**71.77**)	91.46 (57.54)
LFG-UD	90.71 (61.41)	88.86 (56.26)	**93.84** (70.06)	**91.83** (60.09)
UD v1.4	80.58 (38.32)	76.74 (29.04)	91.60 (56.27)	89.26 (45.49)

Table 1: Obtained precision for parsers and treebanks in terms of *Unlabeled Attachment Score* (UAS) and *Labeled Attachment Score* (LAS)

The MATE parser trained on the UD v1.4 treebank performs reasonably well; the scores are comparable to those reported by Øvrelid and Hohle (2016), who give a UAS of 91.21 and a LAS of 88.54 for MATE trained on the UD v1.2 treebank.

They note a significant difference with the scores reported by Solberg et al. (2014) for the Norwegian Dependency Treebank (NDT, the treebank that the Norwegian UD treebanks were derived from), who give a UAS of 92.84 and a LAS of 90.41 for MATE with default training settings and gold-standard POS tags. As they comment, this difference can at least partially be accounted for by the fact that the NDT annotation principles differ from those in UD in some important details: Whereas UD treats prepositions as dependents of the prepositional complement and auxiliaries as dependents of the lexical verb, they are heads in NDT.

We do not see a comparable difference in the performance of the parsers derived from the LFG non-projective dependencies and the UD-style treebanks.

The parser derived from the projective LFG-style dependencies shows a significantly lower LAS than those derived from the non-projective and the UD-style treebanks, which is probably due to the higher number of (non-atomic) relation labels in the projective case.

The scores for the MATE parsers trained on the LFG-derived treebanks may seem to be quite high; these good results should however be viewed critically. The LFG-derived UD treebank and the UD v1.4 treebank are not directly comparable, as the former contains significantly shorter sentences in average, which should make the parse process easier.[17]

It would have been interesting, and perhaps revealing, to test the LFG-derived UD parser on the UD v1.4 treebank test set, but since the treebanks use incompatible POS tags, this cannot be done.

17 The average length of the NorGramBank training sentences is 11.05, with a standard deviation of 5.15, whereas the average length of the UD v1.4 treebank training sentences is 14.81, with standard deviation 8.93.

4 Conclusion

This paper shows how three different types of dependency structures can be derived from LFG analyses, with the c-structure as starting point. All three types of dependency structures can be viewed on the web interface to the XLE parsing framework XLE-Web.[18] When a Norwegian sentence is parsed in XLE-Web, its NorGram LFG analysis is shown, alongside with a dependency structure of a chosen type.

One of the dependency parsers, namely the Stanford NN parser trained on the LFG-proj treebank, has already found a successful application in a quote extraction task (see Salway et al. 2017). In this text mining application, Norwegian newspaper articles were analyzed with the parser, and sentences that contained a speech verb having a politician's name as its subject, or a subject anaphore that could be resolved to a politician's name, were extracted. The sentence complements in the extracted constructions were the desired indirect quotes.

The conversion algorithms described in this paper, in particular the conversion to UD-style dependencies, still need some refinement: some relation types, as well as complex constructions such as comparatives and ellipsis, have not been covered yet.

The implementation of secondary edges and the corresponding conversion to UD enhanced dependencies is also left for future work. No attempt has been made to synchronize the POS tags with Universal POS tags.

5 Acknowledgments

This work and the development of the used resources have been supported by the INESS and CLARINO projects, which received partial funding from the Research Council of Norway. I want to thank the anonymous reviewers for valuable comments and suggestions. Finally, I would like to thank Helge Dyvik for many fruitful discussions, both in the context of this article, and on many other linguistic and non-linguistic topics. Talking with Helge is always a source of inspiration.

References

Bohnet, Bernd (2010). "Very High Accuracy and Fast Dependency Parsing is not a Contradiction". In: *The 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China, pp. 89–97.

Bresnan, Joan (2001). *Lexical-Functional Syntax*. Blackwell Publishers.

Çetinoğlu, Özlem, Jennifer Foster, Joakim Nivre, Deirdre Hogan, Aoife Cahill, and Josef van Genabith (2010). "LFG without C-structures". In: *NEALT Proceedings Series, Vol. 9*, pp. 43–54.

18 http://clarino.uib.no/iness/xle-web

Chen, Danqi and Christopher Manning (2014). "A Fast and Accurate Dependency Parser Using Neural Networks". In: *The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Beijing, China, pp. 740–750.

Choi, Jinho D., Joel Tetreault, and Amanda Stent (2015). "It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Vol. 1. Bejing, China, pp. 387–396.

De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning (2014). "Universal Stanford Dependencies: a Cross-Linguistic Typology". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 4585–4592.

Dyvik, Helge, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse, and Martha Thunes (2016). "NorGramBank: A 'Deep' Treebank for Norwegian". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis. ELRA. Portorož, Slovenia, pp. 3555–3562.

Haug, Dag Tryslew (2012). "From dependency structures to LFG representations". In: *Proceedings of the LFG'12 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Udayana University, Bali, Indonesia, pp. 271–291.

Haug, Dag Tryslew and Marius L. Jøhndal (2012). "Creating a Parallel Treebank of the Old Indo-European Bible Translations". In: *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*. Ed. by Caroline Sporleder and Kiril Ribarov. Marrakech, Morocco, pp. 27–34.

Karlsson, Fred (1990). "Constraint Grammar as a Framework for Parsing Unrestricted Text". In: *Proceedings of the 13th International Conference on Computational Linguistics*. Vol. 3. Helsinki, Finland, pp. 168–173.

Lavelli, Alberto (2016). "Comparing State-of-the-art Dependency Parsers on the Italian Stanford Dependency Treebank". In: *Proceedings CLiC-it 2016 and EVALITA 2016*. Napoli, Italy.

Nivre, Joakim, Johan Hall, and Jens Nilsson (2006). "MaltParser: A Data-Driven Parser-Generator for Dependency Parsing". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (ELREC 2006)*. Genova, Italy.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Sil-

veira, Reut Tsarfaty, and Daniel Zeman (2016). "Universal Dependencies v1: A Multilingual Treebank Collection". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Paris, France: ELRA, pp. 1659–1666.

Øvrelid, Lilja and Petter Hohle (2016). "Universal Dependencies for Norwegian". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis. ELRA. Portorož, Slovenia, pp. 1579–1585.

Øvrelid, Lilja, Jonas Kuhn, and Kathrin Spreyer (2009). "Cross-framework parser stacking for data-driven dependency parsing". In: *TAL* 50.3, pp. 109–138.

Riezler, Stefan and Alexander Vasserman (2004). "Gradient feature testing and l1 regularization for maximum entropy parsing". In: *Proceedings of EMNLP'04*. Barcelona, Spain.

Salway, Andrew, Paul Meurer, and Knut Hofland (2017). "Quote Extraction and Attribution from Norwegian Newspapers". In: *21st Nordic Conference on Computational Linguistics (NoDaLiDa) short papers, forthcoming*. Gøteborg, Sweden.

Solberg, Per Erik, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen (2014). "The Norwegian Dependency Treebank". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (ELREC 2014)*. Reykjavik, Island.

A full-fledged hierarchical lexicon in LFG: the FrameNet approach

Adam Przepiórkowski

Abstract. The aim of this paper is to propose a fully hierarchical organisation of valency information in Lexical Functional Grammar, inspired by recent LFG work on using templates to encode valency. The particular proposal rather closely follows FrameNet's inheritance hierarchy, makes heavy use of templates to encode multiple inheritance, and avoids the problem of multiple inheritance of semantic resources.

1 Introduction

In many constraint-based linguistic theories, as well as in some lexicographic projects, lexical information is organised hierarchically (e.g. Daelemans et al. 1992). In such a hierarchy, internal nodes represent various generalisations pertaining to various portions of the lexicon. These generalisations are inherited by 'lower' nodes. The 'lowest' nodes — the 'leaves' in the hierarchy — typically correspond to specific lexical items, which inherit generalisations from all the nodes on the way up to the root of the hierarchy, and only add truly idiosyncratic information such as the orthographic form. This approach to the lexicon is an important aspect of Head-driven Phrase Structure Grammar (cf., e.g., Flickinger 1987; Davis 2001), but similar proposals have also been made within Lexicalized Tree-Adjoining Grammar (Vijay-Shanker and Schabes 1992) and within Categorial Grammar (Linden 1992), *inter alia*. Hierarchical organisation is also an important feature of WordNet (Fellbaum 1998; Miller et al. 1990) and FrameNet (Fillmore and Baker 2015; Fillmore, Johnson, et al. 2003; Ruppenhofer et al. 2016). While in all these approaches hierarchies represent mainly syntactic and semantic generalisations, Network Morphology (Corbett and Fraser 1993), based on the lexical representation language DATR (Evans and Gazdar 1996), is concerned with morphological and morphosyntactic generalisations.

To the best of my knowledge, the possibility of adopting such a comprehensive taxonomic approach to the lexicon has never been seriously entertained within Lexical Functional Grammar (Bresnan 1982; Bresnan, Asudeh, et al. 2015; Dalrymple 2001). The aim of this paper is to propose an organisation of the LFG lexicon that is close to

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1336.

that of FrameNet. The technical side of this proposal is relatively straightforward, assumes the Glue approach to LFG semantics (Dalrymple 1999, 2001), makes heavy use of templates (Asudeh et al. 2008, 2013; Dalrymple, Kaplan, et al. 2004), and does not require any formal extensions to the underlying LFG machinery, but does require some care to avoid the spurious multiple introduction of meaning constructors. In an accompanying paper (Przepiórkowski 2017a), which shares with the current paper most of the material of the initial three sections, I show that this approach to the lexicon also meshes well with my recent proposal *not* to distinguish arguments from adjuncts in LFG (Przepiórkowski 2016).

2 Inheritance in FrameNet

FrameNet organises lexical knowledge with reference to cognitive structures called *frames*. Various lexical items may evoke the same frame. For instance, the Apply_heat frame is evoked by verbs such as BAKE, FRY, GRILL, STEW, etc. (While FrameNet is concerned with various parts of speech, we only deal with the verbal domain here.) Frames also define *frame elements*, i.e. — simplifying a little — semantic roles which are normally expressed by dependents of lexical items evoking the frame. In the case of Apply_heat, typical frame elements are the Cook and the Food, but also the Container that holds the Food to which heat is applied, the Medium through which heat is applied to Food, etc. In examples (1) and (2), verbs evoking the Apply_heat frame are in boldface.[1]

(1) **Boil** [the potatoes]$_{\text{Food}}$ [in a medium-sized pan]$_{\text{Container}}$.

(2) [Drew]$_{\text{Cook}}$ **sauteed** [the garlic]$_{\text{Food}}$ [in butter]$_{\text{Medium}}$.

Frames are linked via a number of relations, including the hierarchical multiple-inheritance relation. For example, Apply_heat inherits semantic roles from both Activity and Intentionally_affect frames, and the latter inherits from Intentionally_act, which in turn inherits from Event (see Figure 1). It is not clear whether this is a design feature of FrameNet or just a reflection of its work-in-progress status, but it happens in current versions of FrameNet (including the latest at the time of writing this paper, version 1.7) that the same role is introduced multiple times in the hierarchy. For example, within the fragment of the inheritance hierarchy in Figure 1, the Agent role is introduced independently at Activity, at Objective_influence (where it is called Influencing_entity; see below) and at Intentionally_act.

Another feature of FrameNet is that inherited roles, as they acquire more specialised meanings, may change names.[2] For example, the agentive role introduced at Objec-

1 These made up examples are taken from the description of the Apply_heat frame at the FrameNet web interface, at https://framenet2.icsi.berkeley.edu/.

2 This correspondence between frame elements of different frames is currently not shown in the web interface to FrameNet, but it is explicitly defined in the distributed version of the lexicon, in the file frRelation.xml.

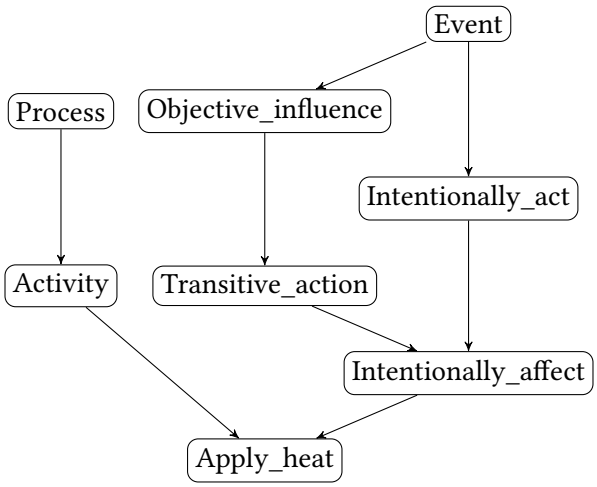


Figure 1: A fragment of the FrameNet 1.7 inheritance hierarchy — all frames from which Apply_heat inherits

tive_influence is actually called Influencing_entity there, but gets renamed to Agent when it is inherited by Transitive_action. The Agent roles of both Transitive_action and Intentionally_act correspond to the (single) Agent role of Intentionally_affect, but the role of Apply_heat corresponding to the Agent roles of both Activity and Intentionally_affect is renamed to Cook. Similarly, the Food role of Apply_heat corresponds to the Patient role of Intentionally_affect and Transitive_action above it (where it is renamed from the Dependent_entity role of Objective_influence). Below, I will simplify by adopting single names for roles related via the inheritance hierarchy. For example, instead of Cook and Food, the respective roles of Apply_heat will be called Agent and Patient, as on the superordinate frames. But, as always, it should be borne in mind that a role on a subordinate frame will usually carry more entailments than the homonymous role on a superordinate frame.

An important aspect of FrameNet is that frame elements correspond to both arguments and adjuncts. For example, among the roles associated with Apply_heat are roles realised by typical adjuncts, such as Manner, Time and Place. A FrameNet reflex of the argument/adjunct dichotomy is its categorisation of roles into core (corresponding to arguments) and non-core (corresponding to adjuncts), but the criteria used for deciding whether a role is core or not suffer from the usual problems (discussed, e.g., in Przepiórkowski 2016) of providing only partial tests or being pairwise incompatible.[3]

What is interesting is that inheritance may change the coreness status of a role. For example, at the Event frame the roles Time and Place are marked as core, proba-

3 See Przepiórkowski (2017a) for further discussion of the core/non-core distinction in FrameNet.

bly reflecting the intuition that verbs directly evoking this frame, such as HAPPEN or OCCUR, seem to require their presence: #*This event occurred* (but this impression is contested below). However, the same roles are treated as non-core on almost all of the 27 frames directly subordinate to Event.[4] The reverse situation happens in the case of the Existence frame, where Time and Place are non-core, but become core on its directly subordinate frame, Circumbscribed_existence. Such changes of coreness seem to bring non-monotonicity to the otherwise monotonic inheritance relation in FrameNet. Below, we will see how such apparently non-monotonic behaviour can be modelled via the monotonic means of Lexical Functional Grammar.

3 Valency in LFG

As is common in LFG, I assume the existence of a level of representation which encodes the semantic argument structure, i.e., which contains information about semantic (or thematic) roles such as Agent or Goal. Traditionally, semantic forms — values of PRED — served this purpose in Lexical Functional Grammar (Kaplan and Bresnan 1982). Alternatively, we could employ the distinct level of argument structure of Butt et al. 1997. Instead, I build here on more recent work and assume the formalisation of argument structure within the semantic structure (Asudeh and Giorgolo 2012; Asudeh, Giorgolo, and Toivonen 2014; Findlay 2016). For example, Asudeh and Giorgolo (2012, p. 78) propose the f-structure and s-structure in Figure 2 for the sentence *Kim tapped Sandy with Excalibur*.

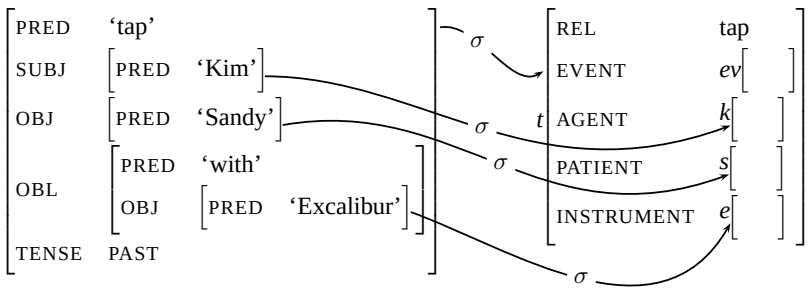


Figure 2: F-structure and s-structure for the sentence *Kim tapped Sandy with Excalibur*

Much of the work in this thread assumes that some of the semantic attributes — those which correspond to arguments in the scope of (Lexical) Mapping Theory (Bresnan and Kanerva 1989) as formalised in Findlay (2016) — are called ARG_1, ARG_2, etc., instead of the more mnemonic names such as AGENT and PATIENT, but — as I am not concerned with LMT in this paper — I will continue to use these more intuitive names.

4 The only exception is the Emergency frame, which treats Time as core.

What kind of lexical entries give rise to such f- and s-structures? Let us consider a simpler case, that of the transitive verb *devoured*; the first version of the lexical entry for this verb is shown in (3).

(3) *devoured* V $(\uparrow \textsc{pred}) = \text{`}\textsc{devour}\text{'}$
$(\uparrow_\sigma \textsc{rel}) = \textsc{devour}$
$\lambda e.\ \mathit{devour}(e) : (\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma$
$(\uparrow \textsc{subj})_\sigma = (\uparrow_\sigma \textsc{agent})$
$\lambda P \lambda x \lambda e.\ P(e) \wedge \mathit{agent}(e, x) :$
$[(\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma] \multimap (\uparrow_\sigma \textsc{agent}) \multimap (\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma$
$(\uparrow \textsc{obj})_\sigma = (\uparrow_\sigma \textsc{patient})$
$\lambda P \lambda x \lambda e.\ P(e) \wedge \mathit{patient}(e, x) :$
$[(\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma] \multimap (\uparrow_\sigma \textsc{patient}) \multimap (\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma$
$(\uparrow \textsc{tense}) = \textsc{past}$
$\lambda P.\ \exists e\ P(e) \wedge \mathit{past}(e) : [(\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma] \multimap \uparrow_\sigma$

There are four natural parts of this lexical entry: 1) the idiosyncratic part, defining PRED,[5] as well as the corresponding s-structure attribute REL,[6] and introducing the basic meaning constructor containing the *devour* relation in its meaning representation; 2) the part saying that the subject grammatical function realises the agent semantic relation; 3) the analogous part defining the correspondence between the object and the patient; 4) the part adding tense information to the f-structure and to the meaning representation, as well as defining the existential closure over the event variable. In the case of the sentence *Godzilla devoured Kim*, this lexical entry gives rise to the f- and s-structures in Figure 3, as well as to the instantiated meaning constructors in (4).

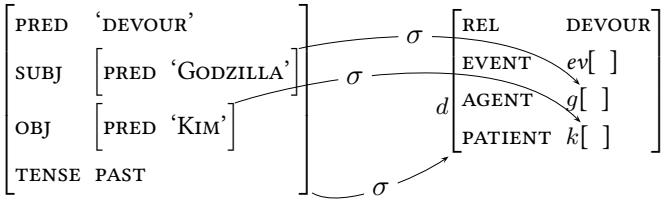


Figure 3: F-structure and s-structure for the sentence *Godzilla devoured Kim*

5 As noted already in Dalrymple, Hinrichs, et al. (1993, pp. 13–14) and Kuhn (2001, § 1.3.3), Glue makes PRED largely superfluous. Here I follow Asudeh and Giorgolo (2012) and retain PRED, but with values reflecting the predicate *sans* its valency — hence, the value of PRED in (3) is defined as 'DEVOUR' rather than 'DEVOUR⟨SUBJ, OBJ⟩'.

6 The existence of the semantic attribute REL is assumed — but not formally introduced — in various recent papers (including Asudeh and Giorgolo 2012; Asudeh et al. 2008, 2013; Asudeh, Giorgolo, and Toivonen 2014; Findlay 2016). The following lexical entries make the introduction of this attribute explicit.

(4)
1. $\lambda e.\ \textit{devour}(e) : \textit{ev} \multimap d$
2. $\lambda P \lambda x \lambda e.\ P(e) \wedge \textit{agent}(e, x) : [\textit{ev} \multimap d] \multimap g \multimap \textit{ev} \multimap d$
3. $\lambda P \lambda x \lambda e.\ P(e) \wedge \textit{patient}(e, x) : [\textit{ev} \multimap d] \multimap k \multimap \textit{ev} \multimap d$
4. $\lambda P.\ \exists e\ P(e) \wedge \textit{past}(e) : [\textit{ev} \multimap d] \multimap d$

These instantiated meaning constructors, together with the instantiated meaning constructors in (5), introduced by the lexical entries of *Godzilla* and *Kim*, may be used to derive the expected meaning representation for the whole sentence: $\exists e\ \textit{devour}(e) \wedge \textit{agent}(e, \textit{godzilla}) \wedge \textit{patient}(e, \textit{kim}) \wedge \textit{past}(e)$.

(5)
5. $\textit{godzilla} : g$
6. $\textit{kim} : k$

One possible proof is shown in (6).

(6)
7. $\lambda x \lambda e.\ \textit{devour}(e) \wedge \textit{agent}(e, x) : g \multimap \textit{ev} \multimap d$ (from 2 and 1)
8. $\lambda e.\ \textit{devour}(e) \wedge \textit{agent}(e, \textit{godzilla}) : \textit{ev} \multimap d$ (from 7 and 5)
9. $\lambda x \lambda e.\ \textit{devour}(e) \wedge \textit{agent}(e, \textit{godzilla}) \wedge \textit{patient}(e, x) : k \multimap \textit{ev} \multimap d$ (from 3 and 8)
10. $\lambda e.\ \textit{devour}(e) \wedge \textit{agent}(e, \textit{godzilla}) \wedge \textit{patient}(e, \textit{kim}) : \textit{ev} \multimap d$ (from 9 and 6)
11. $\exists e\ \textit{devour}(e) \wedge \textit{agent}(e, \textit{godzilla}) \wedge \textit{patient}(e, \textit{kim}) \wedge \textit{past}(e) : d$ (from 4 and 10)

Obviously, apart from the first — idiosyncratic — part of the lexical entry (3), the other three parts will also occur in many other lexical entries, so it makes sense to encode them as templates (Dalrymple, Kaplan, et al. 2004), as in (7)–(9).[7]

(7) $\textsc{agent} := (\uparrow \textsc{subj})_\sigma = (\uparrow_\sigma \textsc{agent})$
$\lambda P \lambda x \lambda e.\ P(e) \wedge \textit{agent}(e, x) :$
$[(\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma] \multimap (\uparrow_\sigma \textsc{agent}) \multimap (\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma$

(8) $\textsc{patient} := (\uparrow \textsc{obj})_\sigma = (\uparrow_\sigma \textsc{patient})$
$\lambda P \lambda x \lambda e.\ P(e) \wedge \textit{patient}(e, x) :$
$[(\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma] \multimap (\uparrow_\sigma \textsc{patient}) \multimap (\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma$

(9) $\textsc{past} := (\uparrow \textsc{tense}) = \textsc{past}$
$\lambda P.\ \exists e\ P(e) \wedge \textit{past}(e) : [(\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma] \multimap \uparrow_\sigma$

7 In the actual templates, the parts defining the correspondence between a semantic argument and a grammatical function will be more complex, to allow for diathesis (see Asudeh, Giorgolo, and Toivonen 2014; Findlay 2016).

With these templates in hand, the lexical entry for *devoured* simplifies to the second (final) version in (10).

(10) *devoured* V $(\uparrow \textsc{pred}) = \text{`}\textsc{devour}\text{'}$
$(\uparrow_\sigma \textsc{rel}) = \textsc{devour}$
$\lambda e.\ \textit{devour}(e) : (\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma$
@AGENT
@PATIENT
@PAST

LFG templates may call other templates, and in this sense they form a kind of hierarchy, although it is a very different kind of hierarchy than, say, the type hierarchy of Head-driven Phrase Structure Grammar (Pollard and Sag 1994). In the context of valency, this possibility was explored by Asudeh et al. (2008, 2013) in their account of English and Swedish *way*-constructions, as in: *Bill elbowed his way through the crowd.* Similarly, Asudeh, Giorgolo, and Toivonen (2014) make use of such embedded template calls for example when defining a prototypical transitive argument structure as in (11).

(11) AGENT-PATIENT := @AGENT @PATIENT

The proposal presented in the following section may be seen as taking the approach summarised above to its logical conclusion.

4 FrameNet-inspired extended valency in LFG

One important difference between LFG work on valency and FrameNet work on semantic roles (or — more generally — frame elements) concerns the argument/adjunct distinction: in LFG valency is understood traditionally, as concerned only with arguments, while FrameNet semantic roles correspond to both arguments and adjuncts. In Section 4.1 we will see that extending the LFG treatment of valency to all dependents — arguments and adjuncts alike — is relatively straightforward. Another important difference is that the idea of the inheritance of valency information has only been applied to a very specific construction in LFG, namely, to the *way*-constructions in a few languages, discussed in Asudeh et al. (2008, 2013), while it is a conspicuous feature of the whole lexicon in FrameNet. In Section 4.2 we will see how this holistic approach of FrameNet may be ported to LFG.

4.1 Frame elements via templates

In LFG, as in most other theories, arguments of a head are selected by this head, i.e., they are logical arguments, while adjuncts are not selected, acting instead as logical functors. In the neo-Davidsonian representations assumed here (Parsons 1990), this distinction is not visible in the final semantic representations, where each dependent — whether an argument or an adjunct — typically introduces a separate predicate (*agent*,

instrument, *beneficiary*, etc.). However, this distinction is present in the grammar and in the lexicon: verifying that a given head may combine with a given dependent is the responsibility of the head when the dependent is an argument, but it is the job of the dependent when the dependent is an adjunct. In practice, there is a lot of LFG work on the first scenario, i.e., on what arguments are required by what heads, but hardly any work on the second scenario, i.e., on what adjuncts are compatible with what heads. The often unspoken assumption is that particular adjuncts, e.g., manner or durative, select only those heads which are semantically compatible, but — to the best of my knowledge — this notion of semantic compatibility has never been formalised or made precise in LFG.

One advantage of FrameNet is that it does model which predicates are compatible with which semantic roles expressed as adjuncts, i.e., it treats adjuncts just as arguments in this respect. For example, referring to the fragment of the inheritance hierarchy in Figure 1, the non-core Duration role is introduced high in the hierarchy, at the Event frame; Means and Purpose are introduced at a subordinate frame, Intentionally_act; Instrument is introduced even lower, at Intentionally_affect; and Medium — also a non-core frame element — is only introduced at a leaf in the hierarchy, at Apply_heat. Hence, in order to implement the FrameNet approach in LFG, templates should be provided not only for typical argument roles, such as AGENT and PATIENT in (7)–(8) above, but also for roles typically realised as adjuncts.

Let us start with *for*-benefactives, which — as discussed for example in Needham and Toivonen (2011, pp. 409–410, 417) — have mixed argument/adjunct properties and should perhaps be analysed as arguments of some predicates and adjuncts of other predicates. For this reason, the analogue of the equation $(\uparrow \textsc{subj})_\sigma = (\uparrow_\sigma \textsc{agent})$ in the above template (7) is the more complex equation $(\uparrow \{\textsc{obl}_{\textsc{ben}} \mid \textsc{adj} \in\})_\sigma = (\uparrow_\sigma \textsc{ben})$, which establishes the correspondence between the semantic attribute BEN(eficiary) and either an argument ($\textsc{obl}_{\textsc{ben}}$) or an adjunct. The definition of the first version of the BENEFICIARY template is given in (12).[8,9]

(12) $$\begin{aligned}\textsc{beneficiary} := \; & (\uparrow \{\textsc{obl}_{\textsc{ben}} \mid \textsc{adj} \in\})_\sigma = (\uparrow_\sigma \textsc{ben}) \\ & \lambda P \lambda x \lambda e.\ P(e) \wedge \textit{beneficiary}(e, x) : \\ & [(\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma] \multimap (\uparrow_\sigma \textsc{ben}) \multimap (\uparrow_\sigma \textsc{event}) \multimap \uparrow_\sigma\end{aligned}$$

8 In fact, a more comprehensive definition of AGENT should also take into consideration two possible realisations — as the subject or as the agentive oblique: $(\uparrow \{\textsc{subj} \mid \textsc{obl}_{\textsc{ag}}\})_\sigma = (\uparrow_\sigma \textsc{agent})$; see Findlay (2016).

9 This template assumes that there is an independent mechanism, such as the traditional PRED coupled with the principles of completeness and coherence, which specifies whether a given predicate combines with an $\textsc{obl}_{\textsc{ben}}$ argument or not. In the current setup, with simpler PRED values and no coherence or completeness (see fn. 5), two different templates would have to be defined: one for benefactive arguments, and another for benefactive adjuncts. This technical inconvenience does not arise on the approach of Przepiórkowski (2016, 2017a).

One piece of information that is missing above is that this template is only concerned with *for*-benefactives. For the sake of concreteness, let us assume that *for* in sentences such as *Kim did it for Sandy* is an asemantic preposition, i.e., that it introduces the attribute PFORM with the value FOR, see (13), and that it is a co-head with the following nominal phrase, see the second disjunct under NP in (14).

(13) *for* P (↑ PFORM) = FOR

(14) PP → P NP
↓=↑ ↓ = (↑ OBJ) | ↓=↑

Then, a modification — using the local name %B — of the template in (12) will do. The second (final) version of the definition of the BENEFICIARY template is given in (15).

(15) BENEFICIARY := %B = (↑ {OBL$_{\text{BEN}}$ | ADJ ∈})
(%B PFORM) =$_c$ FOR
%B$_\sigma$ = (↑$_\sigma$ BEN)
$\lambda P \lambda x \lambda e.\ P(e) \wedge$ *beneficiary*(e, x) :
$[(\uparrow_\sigma$ EVENT) ⊸ $\uparrow_\sigma] \multimap (\uparrow_\sigma$ BEN) ⊸ (↑$_\sigma$ EVENT) ⊸ $\uparrow_\sigma$

This template, just as the templates for AGENT and PATIENT, introduces into the meaning representation a specific predicate, *beneficiary*, and the particular *for*-PP only provides an (e-type) argument for this predicate. The contribution of manner adjuncts, such as *nicely*, is different: it is the role of particular adverbs of manner, rather than the MANNER template, to introduce specific predicates, e.g., *nicely*. Assuming a lexical entry for *nicely* as in (16), the MANNER template may be defined as in (17).[10]

(16) *nicely* Adv (↑ PRED) = 'NICELY'
(↑$_\sigma$ REL) = NICELY
$\lambda e.$ *nicely*(e) : (↑$_\sigma$ REL) ⊸ $\uparrow_\sigma$

(17) MANNER := (↑$_\sigma$ MANNER) = (↑ ADJ ∈)$_\sigma$
$\lambda P \lambda Q \lambda e.\ P(e) \wedge Q(e)$:
$[(\uparrow_\sigma$ EVENT) ⊸ $\uparrow_\sigma] \multimap$
$[(\uparrow_\sigma$ MANNER REL) ⊸ (↑$_\sigma$ MANNER)] ⊸
(↑$_\sigma$ EVENT) ⊸ $\uparrow_\sigma$

An analogous template, shown in (18), is required for locative phrases such as *in Warsaw*.

(18) PLACE := (↑$_\sigma$ PLACE) = (↑ ADJ ∈)$_\sigma$
$\lambda P \lambda Q \lambda e.\ P(e) \wedge Q(e)$:

10 The following constraint may be added to the lexical entry of *nicely* to make sure that it only plays the semantic role of manner: (MANNER ↑$_\sigma$).

$$[(\uparrow_\sigma \ \textsc{event}) \multimap \uparrow_\sigma] \multimap$$
$$[(\uparrow_\sigma \ \textsc{place rel}) \multimap (\uparrow_\sigma \ \textsc{place})] \multimap$$
$$(\uparrow_\sigma \ \textsc{event}) \multimap \uparrow_\sigma$$

The only difference between *nicely* and *in Warsaw* is the kind of relation, Q, provided by the dependent: in the case of *nicely* it was $\lambda e.nicely(e)$, in the case of *in Warsaw* it should be $\lambda e.in(e, warsaw)$. This means that the lexical entry for the semantic preposition *in*, shown in (19), is a little more complex than that for *nicely*, as it must take care of the object of this preposition (here: *Warsaw*).[11]

(19) *in* P $(\uparrow \textsc{pred}) = \text{`}\textsc{in}\text{'}$
$(\uparrow_\sigma \textsc{rel}) = \textsc{in}$
$(\uparrow_\sigma \textsc{loc}) = (\uparrow \textsc{obj})_\sigma$
$\lambda x \lambda e.\ in(e, x) \ : \ (\uparrow_\sigma \textsc{loc}) \multimap (\uparrow_\sigma \textsc{rel}) \multimap \uparrow_\sigma$

Let us take a look at these templates and lexical entries in action, in the sentence *Kim danced nicely for Sandy in Warsaw*. Assuming appropriate syntactic rules, standard lexical entries for proper names, and a lexical entry for *danced* which obligatorily calls the AGENT template and optionally calls the tempates BENEFICIARY, MANNER and PLACE (perhaps among many others), the f-structure and s-structure shown in Figure 4 will result. Moreover, the instantiated meaning constructors in (20) will be added to the

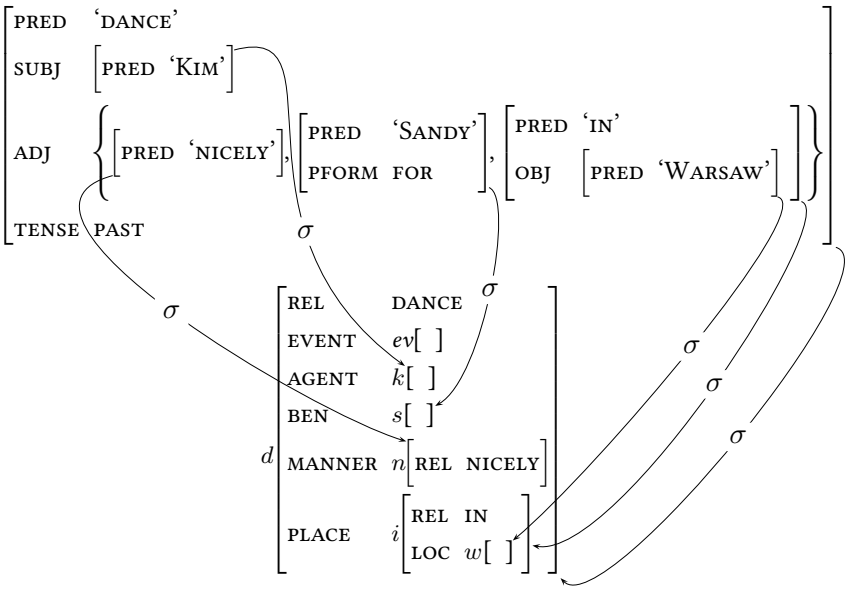


Figure 4: F-structure and s-structure for the sentence *Kim danced nicely for Sandy in Warsaw*

11 Again, the following constraint may be added to the lexical entry for *in* to make sure that the PP based on this lexical entry only plays the semantic role of place: $(\textsc{place} \uparrow_\sigma)$.

pool (1–3 from the lexical entries of proper names, 4 directly from the lexical entry of *danced*, 5 from the PAST template called there, 6–9 from the AGENT, BENEFICIARY, MANNER and PLACE templates called there, 10 from the lexical entry of *nicely* and 11 from the lexical entry of *in*).

(20)
1. $\mathit{kim} : k$
2. $\mathit{sandy} : s$
3. $\mathit{warsaw} : w$
4. $\lambda e.\ \mathit{dance}(e) : \mathit{ev} \multimap d$
5. $\lambda P.\ \exists e\ P(e) \wedge \mathit{past}(e) : [\mathit{ev} \multimap d] \multimap d$
6. $\lambda P \lambda x \lambda e.\ P(e) \wedge \mathit{agent}(e, x) : [\mathit{ev} \multimap d] \multimap k \multimap \mathit{ev} \multimap d$
7. $\lambda P \lambda x \lambda e.\ P(e) \wedge \mathit{beneficiary}(e, x) : [\mathit{ev} \multimap d] \multimap s \multimap \mathit{ev} \multimap d$
8. $\lambda P \lambda Q \lambda e.\ P(e) \wedge Q(e) : [\mathit{ev} \multimap d] \multimap [(n\ \textsc{rel}) \multimap n] \multimap \mathit{ev} \multimap d$
9. $\lambda P \lambda Q \lambda e.\ P(e) \wedge Q(e) : [\mathit{ev} \multimap d] \multimap [(i\ \textsc{rel}) \multimap i] \multimap \mathit{ev} \multimap d$
10. $\lambda e.\ \mathit{nicely}(e)\ :\ (n\ \textsc{rel}) \multimap n$
11. $\lambda x \lambda e.\ \mathit{in}(e, x)\ :\ w \multimap (i\ \textsc{rel}) \multimap i$

It is easy to see that these constructors give rise to the expected meaning representation for this sentence, cf. (21).

(21)
12. $\lambda e.\ \mathit{in}(e, \mathit{warsaw})\ :\ (i\ \textsc{rel}) \multimap i$ (from 11 and 3)
13. $\lambda x \lambda e.\ \mathit{dance}(e) \wedge \mathit{agent}(e, x) : k \multimap \mathit{ev} \multimap d$ (from 6 and 4)
14. $\lambda e.\ \mathit{dance}(e) \wedge \mathit{agent}(e, \mathit{kim}) : \mathit{ev} \multimap d$ (from 13 and 1)
15. $\lambda x \lambda e.\ \mathit{dance}(e) \wedge \mathit{agent}(e, \mathit{kim}) \wedge \mathit{beneficiary}(e, x) : s \multimap \mathit{ev} \multimap d$ (from 7 and 14)
16. $\lambda e.\ \mathit{dance}(e) \wedge \mathit{agent}(e, \mathit{kim}) \wedge \mathit{beneficiary}(e, \mathit{sandy}) : s \multimap \mathit{ev} \multimap d$ (from 15 and 2)
17. $\lambda Q \lambda e.\ \mathit{dance}(e) \wedge \mathit{agent}(e, \mathit{kim}) \wedge \mathit{beneficiary}(e, \mathit{sandy}) \wedge Q(e) :$ $[(n\ \textsc{rel}) \multimap n] \multimap \mathit{ev} \multimap d$ (from 8 and 16)
18. $\lambda e.\ \mathit{dance}(e) \wedge \mathit{agent}(e, \mathit{kim}) \wedge \mathit{beneficiary}(e, \mathit{sandy}) \wedge \mathit{nicely}(e) : \mathit{ev} \multimap d$ (from 17 and 10)
19. $\lambda Q \lambda e.\ \mathit{dance}(e) \wedge \mathit{agent}(e, \mathit{kim}) \wedge \mathit{beneficiary}(e, \mathit{sandy}) \wedge \mathit{nicely}(e) \wedge Q(e) :$ $[(i\ \textsc{rel}) \multimap i] \multimap \mathit{ev} \multimap d$ (from 9 and 18)
20. $\lambda e.\ \mathit{dance}(e) \wedge \mathit{agent}(e, \mathit{kim}) \wedge \mathit{beneficiary}(e, \mathit{sandy}) \wedge \mathit{nicely}(e) \wedge$ $\mathit{in}(e, \mathit{warsaw}) : \mathit{ev} \multimap d$ (from 19 and 12)
21. $\exists e\ \mathit{dance}(e) \wedge \mathit{agent}(e, \mathit{kim}) \wedge \mathit{beneficiary}(e, \mathit{sandy}) \wedge \mathit{nicely}(e) \wedge$ $\mathit{in}(e, \mathit{warsaw}) \wedge \mathit{past}(e) : d$ (from 5 and 20)

In summary, it is possible to define templates introducing various types of dependents, both arguments and adjuncts. In Section 4.2 we will see how to call such templates in a way that does not cause massive redundancy in the resulting description.

4.2 Frame inheritance via template inheritance

The formalisation of FrameNet's multiple-inheritance hierarchy within LFG is relatively straightforward, although some care needs to be taken to avoid multiple introduction of glue resources. I will illustrate such a formalisation with the Apply_heat frame evoked by BOIL.

FrameNet lists 15 semantic roles of the Apply_heat frame. Many of these roles, including Time and Place, but also Agent and Patient, are elements of many different frames. However, an inheritance hierarchy makes it possible to avoid redundancy by introducing particular semantic roles only once or a couple of times in the appropriate place(s) of the hierarchy. Following FrameNet, I assume that the maximally general frame Event introduces the following seven roles potentially realised as dependents of lexical units evoking this frame:[12] Place, Time, Duration, Explanation, Frequency, Manner and Timespan. As mentioned in Section 2, the first two, Place and Time, are marked as core. As also mentioned there, the criteria used to distinguish core and non-core frame elements in FrameNet mirror the vague and pairwise incompatible criteria usually invoked to distinguish arguments from adjuncts. Here, I assume that the intuition behind coreness strongly correlates with the intuition of obligatoriness: the frame elements which are marked as core are usually either syntactically or semantically obligatory in some sense. This also seems to be the reason for marking Place and Time as core: at first sight verbs which directly evoke the Event frame, like HAPPEN or OCCUR, seem to require Time and/or Place. However, this intuition is more naturally explained with a reference to Grice's Maxim of Quantity — arguments similar to those in Goldberg and Ackerman (2001) may be given showing that the requirement of Time or Place is purely pragmatic and may be overriden, as in the following sentence adduced by an anonymous reviewer: *That long-anticipated event didn't occur after all*, or the attested:[13] *Scientists manipulate brains of mice to make them think fake event really occurred*. Obviously, the Event frame pertains to situations which normally occur at some time and at some place, so in this sense Time and Place are semantically obligatory, but the same can be said about all frames inheriting from Event, on which, however, Time and Place are not marked as core. For this reason, I will treat Place and Time as non-core frame elements of Event, just as these roles are treated in FrameNet on frames subordinate to Event. More generally, I am not aware of convincing cases of a role changing its status from core on a superordinate frame to non-core on a subordinate frame, so I will not model this possibility below.

In LFG, frames are naturally encoded as templates which call particular templates corresponding to frame elements. We will assume that templates corresponding to core frame elements are called obligatorily, and those corresponding to non-core

12 I ignore here another type of FrameNet roles, 'core unexpressed', not realisable by dependents.

13 http://www.independent.co.uk/news/science/is-it-inception-total-recall-no-science-fact-false-implanted-in-mice-brains-8732466.html

frame elements are called optionally.[14] In the case of the Event frame, we define the EVENT_FRAME template in (22) which optionally calls templates such as PLACE (defined above), DURATION, etc.

(22) EVENT_FRAME := (@PLACE) (@TIME) (@DURATION) (@EXPLANATION) (@FREQUENCY) (@MANNER) (@TIMESPAN)

An immediately subordinate frame, Intentionally_act, introduces a few additional roles, including the obligatory Agent and the optional Domain, and inherits most of the roles introduced by Event, apart from Timespan.[15] This may be represented via the template in (23).

(23) INTENTIONALLY_ACT_FRAME := @EVENT_FRAME $\neg(\uparrow_\sigma$ TIMESPAN) @AGENT (@DOMAIN) ...

Note that the way Timespan is *not* inherited, via the negative constraint $\neg(\uparrow_\sigma$ TIMESPAN),[16] actually preserves the monotonicity of inheritance, as this semantic role is defined on the higher frame as optional. Similarly, the obligatoriness of a role which is specified as optional on a superordinate frame may be ensured by an existential constraint such as $(\uparrow_\sigma$ MANNER),[17] as in the case of frames evoked by verbs such as BEHAVE, TREAT or WORD, for which the expression of Manner is obligatory.[18]

Another frame, inheriting from Event and adding a few more semantic roles including the obligatory Agent,[19] is Objective_influence, shown in (24), which is a superordi-

14 This is a vast oversimplification, as there is no strong assumption in FrameNet that all core roles are syntactically obligatory; syntactic obligatoriness is just one of a number of partial criteria assumed in FrameNet for deciding whether a role is core or non-core. Moreover, semantic obligatoriness cuts across the class of non-core roles and is the basis of the distinction within this class between peripheral roles (semantically obligatory but — unlike core roles — not central to the meaning of the frame) and extrathematic roles (semantically optional). As discussed in Przepiórkowski (2016, pp. 562–563), dependents may be obligatory in various ways and for various reasons, some of them of pragmatic nature, and I believe that the issue of the proper modelling of these various aspects of obligatoriness requires substantial research.

15 Also Duration is not mentioned in the description of Intentionally_act in the November 2016 release of FrameNet. It is not clear to me whether these omissions are intentional, but I will use the lack of Timespan here to illustrate how such an apparently non-monotonic aspect may be modelled in LFG.

16 As noted by an anonymous reviewer, it would be more elegant not to 'unpack' the TIMESPAN template this way, but rather have a constraint such as ¬@TIMESPAN. However, this would have the effect of negating a sequence — normally interpreted as conjunction — of two statements, one of which is a constructor. As I am not sure what this would mean (but see Przepiórkowski 2017b for a suggestion), I propose this clearer — even if less elegant — encoding.

17 In this case the more elegant obligatory call to the MANNER template would also do (cf. the previous footnote). However, if such obligatory manners are treated as arguments, a separate template would have to be defined (cf. fn. 9).

18 This is not what actually happens in the current release of FrameNet, as there frames evoked by such verbs are not integrated in the inheritance hierarchy.

19 In FrameNet, this role is called Influencing_entity on this frame, but it is renamed to Agent in the subordinate frame Transitive_action.

nate frame of Transitive_action, shown in (25), which in turn introduces the obligatory Patient role.

(24) OBJECTIVE_INFLUENCE_FRAME := @EVENT_FRAME @AGENT
(@CIRCUMSTANCES) ...

(25) TRANSITIVE_ACTION_FRAME := @OBJECTIVE_INFLUENCE_FRAME
@PATIENT ...

Note that, unlike the EVENT_FRAME template, which corresponds to a root frame, the above two templates contain calls to templates corresponding to immediately superordinate frames, thus encoding inheritance.

Both Transitive_action and Intentionally_act are immediately superordinate frames of Intentionally_affect, as formalised in (26).

(26) INTENTIONALLY_AFFECT_FRAME := @INTENTIONALLY_ACT_FRAME
@TRANSITIVE_ACTION_FRAME ...

The technically unfortunate effect of this is that the template call @AGENT is inherited twice.[20] This is a potential problem as this template includes a meaning constructor — see the second and third lines of (7) — so two copies of this constructor will be present whenever the @INTENTIONALLY_AFFECT_FRAME template is called.[21] There is a straightforward solution to this problem, though, consisting in the following modification of the AGENT template shown in (27), to be compared with the previous definition in (7).

20 Also various other templates are inherited twice, but since they are optional in the first place, this will not lead to the problem discussed here.

21 An anonymous reviewer contests the view that this is a potential problem, saying that "the problem with multiple introduction of glue resources seems to come from a confusion between the mechanism (templates) that express generalizations over entries for lexical and morphological formatives, the structures that those entries describe, and the operations that apply to those structures. Inheritance by template invocation in LFG just gives pieces of text that are used to specify entries that then enter into the grammar and Glue interpreters. I don't know that there is an assumption anywhere that 2 copies of the same text in a lexical scope leads to different behavior than a single instance, even if the entry ultimately has resource sensitive components. The confusion is between the template mechanism for specifying lexical formatives (which basically operates by manipulation and substitution of text strings) and whatever interpretation (like glue deductions) applies to the so-specified formatives. @AGENT @AGENT (or any other stutter) should be the same as a single @AGENT in the specification how to create a formative before it enters into the combinatorial operations of the grammar or semantics…" If this is right, than the slight complication introduced below is not necessary. However, as LFG lacks a comprehensive mathematical formalisation comparable to the formalisation of Head-driven Phrase Structure Grammar (as provided in Richter 2000), it is not clear to me whether a repeated call to a template containing a semantic resource should be interpreted as resulting in one or multiple copies of the resource, and for this reason I provide a solution for the worst-case scenario. This and other questions about the formal status of meaning constructors and semantic structures are addressed in Przepiórkowski (2017b).

(27) AGENT := ($(\uparrow \text{SUBJ})_\sigma = (\uparrow_\sigma \text{AGENT})$
$\lambda P \lambda x \lambda e.\ P(e) \wedge \mathit{agent}(e, x)$:
$[(\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma] \multimap (\uparrow_\sigma \text{AGENT}) \multimap (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma$)
$(\uparrow_\sigma \text{AGENT})$

In the above template, the content of the previous version of this template is made jointly optional (see the parentheses in bold), and a non-optional constraint is added ensuring that the semantic AGENT feature is defined. Note that multiple occurrences of the $(\uparrow_\sigma \text{AGENT})$ constraint have exactly the same effect as a single occurrence, so multiple inheritance of this part of the template is not harmful. The only place in the grammar that assigns a value to the AGENT feature is the optional part of the new AGENT template in (27), so — in case of multiple calls to this template — at least one copy of this optional part must actually be used. On the other hand, at most one may be used, as more would introduce multiple copies of the meaning constructor within the optional part. Each such constructor causes a consumption of the glue resource introduced by the subject, corresponding to the value of the AGENT attribute, and only one such resource is introduced by the subject. Hence, exactly one of the multiple copies of the optional part of the AGENT template will actually be used.

Returning to the running example, the Apply_heat frame in (28) also inherits from two superordinate frames, Intentionally_affect and Activity (the latter not discussed here), and also introduces a few specific roles such as Container and Medium.

(28) APPLY_HEAT_FRAME := @INTENTIONALLY_AFFECT_FRAME
@ACTIVITY_FRAME (@CONTAINER) (@MEDIUM) ...

With such a hierarchy of templates, the lexical entry for *boiled*, introducing the many possible dependents of this verb, boils down to (29).

(29) *boiled* V $(\uparrow \text{PRED}) = \text{'BOIL'}$
$\lambda e.\ \mathit{boil}(e) : (\uparrow_\sigma \text{EVENT}) \multimap \uparrow_\sigma$
@APPLY_HEAT_FRAME
@PAST

In practice, different verbs belonging to the same frame may additionally introduce specific — possibly different — morphosyntactic constraints on the realisation of the same role, as is the case with the verbs BEGIN and ENTER evoking the Activity_start frame: only BEGIN may realise the Activity role as an infinitival phrase (*begin to negotiate*) and only ENTER may realise it as an *into*-PP (*enter into negotiations*).

5 Conclusion

In this paper I proposed to carry over the main ideas of FrameNet to LFG: introduce all kinds of dependents lexically (which does not preclude constructional analyses like

that of Asudeh et al. 2013), and organise such extended valency information hierarchically, so as to avoid redundancy and capture generalisations. This proposal may be seen as pushing to the limit some of the ideas presented in Asudeh and Giorgolo (2012), Asudeh, Giorgolo, and Toivonen (2014) and Asudeh et al. (2008, 2013). While the present paper proposes a way to encode a FrameNet-like hierarchical valency lexicon in standard LFG, an accompanying paper (Przepiórkowski 2017a) shows that this approach to valency meshes particularly well with my earlier proposal *not* to distinguish arguments from adjuncts in LFG (Przepiórkowski 2016) and helps remove the last vestiges of the ill-defined argument/adjunct distinction from this linguistic framework.

Acknowledgments

This paper is dedicated to Helge Dyvik, who is obviously no stranger not only to LFG and formal semantics, but also to hierarchical lexicons (Dyvik 2004). Many thanks are due to the three anonymous reviewers of this volume, whose comments have led to numerous improvements; I also benefitted from comments by Michael Ellsworth, Jamie Y. Findlay, and the audience of the 23rd South of England LFG meeting (in May 2017). The research reported here is partially supported by the Polish Ministry of Science and Higher Education within the CLARIN ERIC programme 2016–2018 (http://clarin.eu/).

References

Asudeh, Ash, Mary Dalrymple, and Ida Toivonen (2008). "Constructions with Lexical Integrity: Templates as the Lexicon–Syntax Interface". In: *The Proceedings of the LFG'08 Conference*. Ed. by Miriam Butt and Tracy Holloway King. University of Sydney, Australia: CSLI Publications, pp. 68–88.

– (2013). "Constructions with Lexical Integrity". In: *Journal of Language Modelling* 1.1, pp. 1–54.

Asudeh, Ash and Gianluca Giorgolo (2012). "Flexible Composition for Optional and Derived Arguments". In: *The Proceedings of the LFG'12 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford, CA: CSLI Publications, pp. 64–84.

Asudeh, Ash, Gianluca Giorgolo, and Ida Toivonen (2014). "Meaning and Valency". In: *The Proceedings of the LFG'14 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford, CA: CSLI Publications, pp. 68–88.

Bresnan, Joan, ed. (1982). *The Mental Representation of Grammatical Relations*. MIT Press Series on Cognitive Theory and Mental Representation. Cambridge, MA: The MIT Press.

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler (2015). *Lexical-Functional Syntax*. 2nd ed. Blackwell Textbooks in Linguistics. Wiley-Blackwell.

Bresnan, Joan and Jonni M. Kanerva (1989). "Locative Inversion in Chichewa: A Case Study of Factorization in Grammar". In: *Linguistic Inquiry* 20.1, pp. 1–50.

Butt, Miriam, Mary Dalrymple, and Anette Frank (1997). "An Architecture for Linking Theory in LFG". In: *The Proceedings of the LFG'97 Conference*. Ed. by Miriam Butt and Tracy Holloway King. University of California, San Diego: CSLI Publications.

Corbett, Greville G. and Norman M. Fraser (1993). "Network Morphology: a DATR Account of Russian Nominal Inflection". In: *Journal of Linguistics* 29, pp. 113–142.

Daelemans, Walter, Koenraad De Smedt, and Gerald Gazdar (1992). "Inhertance in Natural Language Processing". In: *Computational Linguistics* 18.2. Ed. by Walter Daelemans and Gerald Gazdar, pp. 205–218.

Dalrymple, Mary, ed. (1999). *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*. Cambridge, MA: The MIT Press.

– (2001). *Lexical Functional Grammar*. San Diego, CA: Academic Press.

Dalrymple, Mary, Angie Hinrichs, John Lamping, and Vijay Saraswat (1993). "The Resource Logic of Complex Predicate Interpretation". In: *Proceedings of ROCLING 1993*, pp. 3–21.

Dalrymple, Mary, Ronald M. Kaplan, and Tracy Holloway King (2004). "Linguistic Generalizations over Descriptions". In: *The Proceedings of the LFG'04 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford, CA: CSLI Publications, pp. 199–208.

Davis, Anthony R. (2001). *Linking by Types in the Hierarchical Lexicon*. Stanford, CA: CSLI Publications.

Dyvik, Helge (2004). "Translations as Semantic Mirrors: From Parallel Corpus to Wordnet". In: *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23), Göteborg, 22–26 May 2002*. Ed. by Karin Aijmer and Bengt Altenberg. Vol. 49. Language and Computers, pp. 311–326.

Evans, Roger and Gerald Gazdar (1996). "DATR: A Language for Lexical Knowledge Representation". In: *Computational Linguistics* 22.2, pp. 167–216.

Fellbaum, Christiane, ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Fillmore, Charles J. and Collin Baker (2015). "A Frames Approach to Semantic Analysis". In: *The Oxford Handbook of Linguistic Analysis*. Ed. by Bernd Heine and Heiko Narrog. 2nd. Oxford: Oxford University Press, pp. 791–816.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R.L. Petruck (2003). "Background to FrameNet". In: *International Journal of Lexicography* 16.3, pp. 235–250.

Findlay, Jamie Y. (2016). "Mapping Theory without Argument Structure". In: *Journal of Language Modelling* 4.2, pp. 245–289.

Flickinger, Daniel (1987). "Lexical Rules in the Hierarchical Lexicon". Ph.D. Thesis. Stanford, CA: Stanford University.

Goldberg, Adele E. and Farrell Ackerman (2001). "The Pragmatics of Obligatory Adjuncts". In: *Language* 77.4, pp. 798–814.

Kaplan, Ronald M. and Joan Bresnan (1982). "Lexical-Functional Grammar: A Formal System for Grammatical Representation". In: *The Mental Representation of Grammatical Relations*. Ed. by Joan Bresnan. MIT Press Series on Cognitive Theory and Mental Representation. Cambridge, MA: The MIT Press, pp. 173–281.

Kuhn, Jonas (2001). "Resource Sensitivity in the Syntax-Semantics Interface: Evidence from the German Split NP Construction". In: *Constraint-Based Approaches to Germanic Syntax*. Ed. by Detmar Meurers and Tibor Kiss. Stanford, CA: CSLI Publications, pp. 177–215.

Linden, Erik-Jan van der (1992). "Incremental Processing and the Hierarchical Lexicon". In: *Computational Linguistics* 18.2, pp. 219–238.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller (1990). "Introduction to WordNet: An Online Lexical Database". In: *International Journal of Lexicography* 3.4, pp. 235–244.

Needham, Stephanie and Ida Toivonen (2011). "Derived Arguments". In: *The Proceedings of the LFG'11 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford, CA: CSLI Publications, pp. 401–421.

Parsons, Terence (1990). *Events in the Semantics of English: A Study in Subatomic Semantics*. Cambridge, MA: The MIT Press.

Pollard, Carl and Ivan A. Sag (1994). *Head-driven Phrase Structure Grammar*. Chicago, IL: Chicago University Press / CSLI Publications.

Przepiórkowski, Adam (2016). "How *not* to Distinguish Arguments from Adjuncts in LFG". In: *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*. Ed. by Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King, and Stefan Müller. Stanford, CA: CSLI Publications, pp. 560–580.

– (2017a). "Hierarchical Lexicon and the Argument/Adjunct Distinction". Submitted to LFG 2017 proceedings.

– (2017b). "Some Doubts about Meaning Constructors and Semantic Structures in LFG + Glue". Unpublished manuscript.

Richter, Frank (2000). "A Mathematical Formalism for Linguistic Theories with an Application in Head-Driven Phrase Structure Grammar". Ph.D. Thesis. Universität Tübingen.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk (2016). *FrameNet II: Extended Theory and Practice*. Revised November 1, 2016.

Vijay-Shanker, K. and Yves Schabes (1992). "Structure Sharing in Lexicalized Tree-Adjoining Grammars". In: *Proceedings of the 14th International Conference on Computational Linguistics (COLING 1992)*. Nantes, pp. 205–211.

Norwegian bare singulars revisited

Victoria Rosén and Kaja Borthen

Abstract. Borthen (2003) analyzed bare singulars in Norwegian. In this paper some of the claims made there are reexamined by searching in NorGramBank. The study provides new empirical support to the claims put forth in the earlier work, but it also reveals problems with the prior analysis.

1 Introduction

A bare singular is a countable, singular and indefinite nominal constituent that does not have a phonetically realized determiner. Whereas some languages do not have indefinite articles at all, other languages do while still allowing for bare singulars in some cases. These languages include Danish (Asudeh and Mikkelsen 2000), Swedish (Delsing 1993), English (Stvan 1998), Dutch, German and French (De Swart and Zwarts 2009), Hungarian (Kiefer 1994), Albanian (Kallulli 1999), and Brazilian Portuguese (Schmitt and Munn 1999), to mention a few.

In spite of crosslinguistic similarities, the distribution pattern of bare singulars in languages that have indefinite articles varies. For instance, Norwegian allows for bare singulars more extensively than English does. Examples of bare singulars in Norwegian are shown in (1)–(4). Bare singular NPs are marked with boldface in examples here and throughout the article.

(1) *Per er* ***lærer***.
Per is teacher
'Per is a teacher.'

(2) *Han kjører* ***bil***.
he drives car
'He drives a car.'

(3) ***Hest*** *er et koselig dyr.*
horse is a nice animal
'The horse is a nice animal.'

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1330.

(4) ***Taxi*** *er* *dyrt.*
taxi is expensive
'Taking a taxi is expensive.'

Norwegian bare singulars must often be translated into English with an indefinite article, as in (1) and (2). Sometimes however, a definite article is more appropriate, as in the generic statement in (3). In still other cases, neither type of determiner will suffice, and the translation must be rendered periphrastically, as in (4).

English has a quite restrictive use of bare singulars (see e.g. Stvan 1998). According to De Swart and Zwarts (2009), there are five constructions that license bare singulars in English, named 'bare location', 'bare coordination', 'bare predication', 'bare reduplication' and 'bare incorporation'. Their examples (ibid. p. 280) are given in (5)–(9).

(5) John is in **hospital**. (Bare location)

(6) the way to use **knife** and **fork** (Bare coordination)

(7) Mary is **chair** of the department. (Bare predication)

(8) He found **door** after **door** closed. (Bare reduplication)

(9) She is playing **piano** for the choir. (Bare incorporation)

The meanings expressed in (5)–(9) are possible with bare singulars in Norwegian, and more generally, bare singulars in Norwegian and English share important features. Still, the use of bare singulars in each language shows idiosyncratic patterns. As De Swart and Zwarts (2009, p. 7) put it, bare singulars operate "[…] at the border […] of syntax and lexicon, of rules and lists, of regularities and idioms". This poses a particularly strong need for thorough empirical investigations, both in order to accurately describe the distribution pattern of bare singulars in one particular language, and in order to detect crosslinguistic similarities and differences.

In her PhD thesis, Borthen (2003) provided an analysis of Norwegian bare singulars, attempting to account for the necessary and sufficient conditions for the use of these phrases. At the time the research was begun, in 1999, it was not straightforward to search for bare singulars in electronic corpora. In the first place, such corpora for Norwegian were not annotated for the distinction between mass and count nouns. In the second place, it was complicated to pick out NPs without determiners from existing corpora. Several searches for consecutive words such as [V + $N_{sg,indef}$], [V + Adj + $N_{sg,indef}$], and [P + $N_{sg,indef}$] were used. However, such search expressions cannot reliably identify bare singulars, and there is also the risk of finding them only in certain syntactic positions. It was furthermore practically impossible to search for bare singulars in specific constructions. Since it was difficult to reliably find bare singulars

through corpus searches, the data used for the thesis comprised approximately 400 manually collected examples from a number of texts plus a large number of examples that were invented by the author.

The availability of treebanks, which are syntactically annotated corpora, marks a radical change in the study of language. NorGramBank (Dyvik et al. 2016) is a treebank of modern Norwegian constructed by automatically parsing a corpus with NorGram, a computational LFG grammar for Norwegian (Dyvik 2000). A small part of the corpus (approx. 315,000 words) was manually disambiguated using computer-generated discriminants, while the rest (approx. 60 million words) was stochastically disambiguated. NorGramBank was developed in the INESS[1] treebanking infrastructure project (Rosén et al. 2012), which also developed the search language INESS Search (Meurer 2012). The detailed syntactic annotation in NorGramBank and the sophisticated search language make it possible to conduct very fine-grained searches for exactly the phenomena the researcher is interested in.

The main goal of this study is to test some aspects of the theoretical analysis of Borthen (2003) on new, authentic data. We claim that the data made available through NorGramBank and the search options provided by INESS Search constitute an excellent basis for improving the theoretical analysis of the phenomenon under investigation.

2 Bare singulars in Norwegian

Borthen (2003) makes a number of observations regarding the syntactic properties of Norwegian bare singulars, some of which are listed below (ibid. p. 68).

Syntactic properties of Norwegian bare singulars:

- They can occur in all basic syntactic positions available for nominal phrases in Norwegian, but not "freely".
- They can be modified and coordinated.
- They are usually not affected by syntactic alternations such as nominalization, passivization, topicalization, raising, question formation, and subject-object alternations for arguments of presentational verbs.
- Adverbs can freely intervene between Norwegian bare singulars and their co-occurring verbal predicates.

That Norwegian bare singulars cannot occur "freely" in nominal positions means, for instance, that they often occur as direct objects, but not of just any verb and not in just any context. Similarly, they sometimes occur as subjects, but only rarely. This means that the generation of bare singulars is not as productive as that of singulars with overt determiners. On the other hand, we cannot account for Norwegian bare

1 http://clarino.uib.no/iness

singulars merely by assuming that they are part of fixed multiword expressions. Such an explanation is unlikely since bare singulars can be modified and coordinated, adverbs can freely intervene between bare singulars and their selecting predicates, and their acceptability is usually not affected by syntactic alternations. As we will see later in this paper, additional evidence for the view that the generation of Norwegian bare singulars is productive comes from the high number of unique combinations of verbs and bare singulars.

Four construction types were posited in Borthen (2003) in order to predict the productive use of Norwegian bare singulars. Each construction type is illustrated with some examples below (ibid. p. 117, 165, 171, 194, 212, 215).

The 'conventional situation type' construction

(10) *Hun er **elev**.*
she is pupil
'She is a pupil.'

(11) *Hun går på **skole**.*
she goes to school
'She goes to school.'

The 'profiled *have*-predicate' construction

(12) *Hun hadde **rød ytterfrakk**.*
she had red coat
'She had a red coat.'

(13) *Han mangler **sovepose** og **regnfrakk**.*
he lacks sleeping bag and rain coat
'He lacks a sleeping bag and a rain coat.'

(14) *Vi trenger **nytt telt**.*
we need new tent
'We need a new tent.'

(15) *et bord med **hvit duk***
a table with white cloth
'a table with a white cloth'

(16) *Hva skulle vi gjort uten **do**?*
what should we done without toilet
'What should we have done without a toilet?'

The 'taxonomic' construction

(17) *Det hjelpemiddelet som er mest brukt er* ***datamaskin****.*
the tool that is most used is computer
'The type of tool that is used the most is the computer.'

(18) ***Buss*** *er et naturvennlig kjøretøy.*
bus is a nature friendly vehicle
'A bus is a non-polluting vehicle.'

The 'covert infinitival clause' construction

(19) ***Sykkel*** *er kult.*
bike is cool
'To ride a bike is cool.'

(20) *Jeg vil anbefale* ***telt****.*
I will recommend tent
'I would recommend (having/using) a tent.'

According to Borthen (2003, p. 153–154), the 'conventional situation type' construction licenses bare singulars as long as the bare singular and its selecting predicate denote a conventional situation type. A conventional situation type is a property, state, or activity type that occurs frequently or standardly in a given contextual frame and has particular relevance in this frame as a recurring situation type (Borthen 2003, p. 160). This predicts that bare singulars such as *er elev* 'is a pupil' in (10) and *går på skole* 'goes to school' in (11) will be acceptable as long as the verb phrases they are part of are intended to describe a conventional situation type. This construction is more general than the ones that license bare singulars in English, and it subsumes the constructions called 'bare location' (5), 'bare predication' (7), and 'bare incorporation' (9) in English. The constructions that license bare singulars may lead to the development of multiword expressions with bare singulars over time, and the 'conventional situation type' construction is particularly likely to do so because of phonological and semantic characteristics of the construction (Borthen 2003, p. 153–154).

The 'profiled *have*-predicate' construction licenses bare singular arguments on certain interpretations of *have*-predicates, according to Borthen (2003). A *have*-predicate is a predicate that expresses a *have*-relation (an asymmetrical coexistence relation) directly or that can be decomposed into a structure that includes one. For instance, to lack something means to not have something, and to need something means to have a desire or urge to have something. Thus, the verbs *mangle* 'order' and *trenge* 'need' in (13) and (14) are *have*-predicates. Similarly, the preposition *med* 'with' in (15) can

be seen as denoting a *have*-relation directly whereas the preposition *uten* 'without' in (16) denotes a negated *have*-relation. Bare singulars are licensed as arguments of *have*-predicates as long as the context is such that the focus is on the state in which the denotation of the bare singular simply coexists with some other entity mentioned in the sentence (Borthen 2003, p. 187–188). Due to this construction, verbs such as *ha* 'have', *ønske seg* 'wish for', *mangle* 'lack', *få tak i* 'get hold of', *ta med* 'bring', *hente (seg)* 'fetch (for oneself)', and *ta med (seg)* 'bring (for oneself)' allow for bare singular objects on certain interpretations in Norwegian.

The third bare singular licensing construction was originally called the 'comparison of types' construction in Borthen (2003). It has been renamed here as the 'taxonomic' construction, due to the fact that the denotation of the bare singular is presented as having a specific position in a taxonomic hierarchy. Illustrated in (17) and (18), this construction always involves the copular verb *være* 'be' and one preverbal and one postverbal nominal phrase where the bare singular is presented as a hyponym of the denotation of the other nominal phrase. This construction often licenses bare singular subjects.

Finally, according to Borthen (2003), Norwegian has a construction which allows for a "covert infinitival clause interpretation" of indefinite noun phrases in subject or object position. This construction is, however, not a construction that licenses bare singulars directly; it licenses them only if the underlying predication (in the covert infinitival clause) is one which would naturally be expressed by a phrase containing a bare singular object (ibid. p. 222). Thus, bare singulars that occur in this kind of example, as in (19) and (20), could be considered to be licensed by the 'conventional situation type' construction or the 'profiled *have*-predicate' construction.

As for why bare singulars are licensed by the four constructions listed above, Borthen (2003) argues that this has to do with their semantics. Some semantic characteristics of Norwegian bare singulars are listed below (ibid. p. 50–51).

Semantic properties of Norwegian bare singulars:

- They can never take wide scope.
- They can never be referential.
- They can never be partitive.
- They can be generic, but not with a (quasi-)universal interpretation.
- They are poorer antecedent candidates of token pronouns than corresponding expressions with indefinite articles, but they can be antecedents of some identity-of-sense anaphors.
- Their descriptive content cannot be too general.

In order to account for these properties and the construction types that license bare singulars, Borthen (2003) assumes that bare singulars are *type emphasizing*. All countable nouns have a dual aspect to them; on the one hand they denote a property, a

type of thing, and on the other hand they may be used to refer to tokens in the world. Whereas indefinites with the indefinite article indicate relative emphasis on the token involved in the given situation, bare singulars emphasize the type of thing introduced and are only licensed in specific constructions that go naturally along with such interpretations. The syntactic constructions that allow for bare singulars are thus motivated but not fully predicted by the semantics of bare singulars.

As mentioned earlier, bare singulars in languages that have the indefinite article share many properties. Still, there are crosslinguistic differences. Borthen (2003) explains this by proposing that type emphasis is a scalar notion. That is, bare singulars in various languages may point to different positions on a scale of type/token emphasis. Bare singulars across languages are similar because they are all type emphasizing (compared to corresponding phrases with the indefinite article). As such, they are restricted semantically as well as destined to appear in constructions that go particularly well together with type emphasis. At the same time, bare singulars are different across languages since they may differ with respect to where on the scale of type emphasis they are positioned. This, in turn, affects the set of constructions that license them (Borthen 2003, p. 226–227).

3 Problems with Borthen (2003)

Many of the claims put forth in Borthen (2003) are based on invented examples, introspection and impressions. For instance, the following statement is made: "*Ønske seg* 'want' and *dele ut* 'hand out' belong to a semantically related group of verbs that co-occur particularly easily, and thus relatively frequently, with bare singulars in Norwegian" (Borthen 2003, p. 164). The group of verbs referred to in this quote are the ones labeled *have*-predicates above. The claim that bare singulars licensed by these verbs are particularly frequent would be more convincing if it were supported by authentic examples and, for the frequency claim, some statistics.

Another claim in Borthen (2003) is that bare singulars "[…] tend to be unacceptable if they have too little descriptive content"; the invented examples in (21) and (22) are meant to illustrate this (ibid. p. 50).

(21) *Det ligger **kniv** på bordet.*
it lies knife on the table
'There is a knife on the table.'

(22) *??Det ligger **ting/dings/greie** på bordet.*
it lies thing/gizmo/thingamajig on the table
'There is a thing/gizmo/thingamajig on the table.'

Whereas the example in (21) is perfectly fine, the examples in (22) are intuitively unnatural. This claim was based on the intuitions of the author, and would be more convincing with empirical evidence to back it up.

The original study also claimed that "[…] the extensive use of bare singulars in idioms and as part of multi-word lexical entries is striking. In fact, they seem to be more frequent than bare singulars licensed by the general constructions proposed in this thesis" (ibid. p. 342). The claim was also made that "Norwegian bare singulars are usually not affected by syntactic alternations like nominalization, passivization, topicalization, […]" (ibid. p. 68), from which it follows that bare singulars should show up in these sentence structures also in authentic language use. Again, the question must be asked what evidence there is for these claims.

4 Searching for evidence in NorGramBank

We conducted searches in NorGramBank for evidence that can answer research questions such as the following, posed in Borthen (2003).

1. Is it true that bare singular nouns with only very general descriptive content such as *ting* 'thing', *dings* 'gizmo' and *greie* 'thingamajig' are particularly unlikely?

2. What verbs tend to take bare singular arguments? Is it true that *have*-predicates are particularly frequent?

3. What are the most common verb–noun combinations? Is it true that there are more instances of idiomatic expressions with bare singulars than productive uses?

4. Can bare singulars occur in all kinds of non-canonical sentence structures, for example in topicalizations, left-dislocations and clefts?

In 4.2–4.5 these questions will be examined based on searches in NorGramBank.

4.1 Bare singulars in NorGramBank

In order to find bare singulars in NorGramBank, we must know what characterizes them. In addition, we must know how bare singulars are represented in NorGramBank to know what features of the treebank annotation to search for.

A bare singular noun phrase is headed by a count noun in the singular form and does not have an article or a determiner. In English and many other languages, this is a sufficient characterization. In Norwegian, however, we must add that the phrase must be indefinite. The reason for this is that Norwegian nouns are inflected for definiteness; in (23) the noun *lærer* 'teacher' is inflected for definiteness by adding the singular definite suffix *-en* to the stem.

(23) *Per er* ***læreren***.
Per is the teacher
'Per is the teacher.'

The lack of a determiner is therefore not a sufficient criterion for the phrase being indefinite; the noun must also be in the indefinite form.

The syntactic annotation in NorGramBank is in the Lexical-Functional Grammar (LFG) formalism (Bresnan 2001; Dalrymple 2001). Each sentence has a constituent structure (c-structure) and a functional structure (f-structure). The c-structure is a context-free phrase structure tree showing the relations of dominance and linear precedence. The f-structure is an attribute–value matrix which provides information about syntactic functions, such as subject and object, and grammatical features, such as number, gender and tense. The properties that characterize bare singulars are represented in the f-structure. We can examine some f-structures to see how this is done. Figure 1 shows the f-structure for the noun *lærer* 'teacher'.

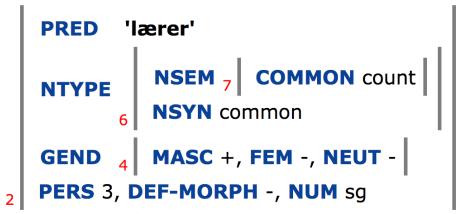


Figure 1: F-structure for the noun *lærer*

The f-structure consists of unordered pairs of attributes and values. Some attributes have simple values; an example is the attribute NUM (number) which has the value sg. Some attributes, such as NTYPE, have other f-structures as their values. The value of NTYPE is a new f-structure (labeled with the index '6') which has an attribute NSEM, which in turn has an f-structure (labeled with the index '7') as its value. The innermost f-structure has the attribute COMMON with the value count. In such cases we speak of a *path* of attributes that leads to a value; here the path is NTYPE NSEM COMMON. The attribute PRED (for *predicate*) has a special type of value called a *semantic form*. This is usually the citation form of the word in single quotes, sometimes followed by a list of arguments, such as for the subcategorized arguments of verbs. The PRED value in the f-structure in Figure 1 is 'lærer'.

For a noun phrase to be a bare singular, its f-structure must have the values sg and count. In addition, the noun phrase must not have a determiner and it must not be definite. In Figure 1 there is an attribute DEF-MORPH with the value – (minus). This means that the noun is in the indefinite form, but not necessarily that the noun phrase is indefinite, since an indefinite noun may occur together with a definite determiner in a definite noun phrase. We therefore need to know that the noun phrase is not definite and that the noun phrase does not have a determiner.

In Figure 2 the f-structure for the noun phrase *denne læreren* 'this teacher' is shown. This noun phrase has an attribute DEF with the value +; this is the value that must not

be present in order for the noun phrase to be indefinite. The determiner *denne* 'this' is represented by the f-structure with the attribute SPEC and its value; the attribute SPEC does not occur in the f-structure of a bare singular.

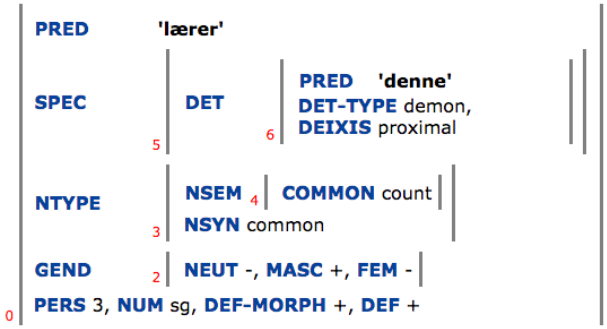


Figure 2: F-structure for the noun phrase *denne læreren*

Searching in NorGramBank is done with INESS Search. The search expression in (24) is designed to pick out bare singular nouns.

(24) #x_ >PRED #p &
#x_ >NUM 'sg' &
#x_ >(NTYPE NSEM COMMON) 'count' &
!(#x_ >SPEC) &
!(#x_ >DEF '\+') &
!(#x_ >PRED 'pro') &
!(#x_ >(OBL PSEM) 'part')

This expression searches for properties in the f-structure. It consists of seven conjuncts, each of which constrains the search to certain properties which the analysis either must have or must not have. The first conjunct says that there is an f-structure #x_ that has an attribute PRED with the value #p.[2] The second conjunct states that #x_ has an attribute NUM with the value 'sg' (atomic f-structure values must be enclosed in single quotes in INESS Search expressions). The third conjunct says that #x_ has a path of attributes NTYPE NSEM COMMON with the value 'count'. The exclamation point in the final four conjuncts is the negation operator; these conjuncts state which properties the f-structure must not have. It must not have a SPEC, it must not have a definite marking, and it must not have the value 'pro' for its PRED. The final conjunct states that the f-structure must not have a path of attributes OBL PSEM with the value 'part' for *partitive*; this ensures that the lexical item that is the value of #x_ is not a

2 All node variables in INESS Search expressions are marked with either #, in which case the node variable is taken to be existentially quantified, or with %, in which case it is universally quantified.

quantifier in a partitive phrase. Together these properties target bare singulars as they are represented in NorGramBank.

INESS Search can present search results as frequency tables, making it easy to examine the results. The search expression indicates which elements are to be displayed in the table by the use of an underscore. If a variable has an underscore, its values are not shown in the table; if a variable does not have an underscore, its values are shown together with their frequencies.

Figure 3 shows the c- and f-structures for (25), one of the sentences found by the search expression in (24). INESS Search highlights the parts of the structure that were searched for in the results. The f-structure of the NP headed by the noun *natt* 'night' is marked by a red box and labeled in the top left corner by the variable #x_. The values required by the search expression are also marked by red boxes.

(25) *Det var **natt**.*
it was night
'It was nighttime.'

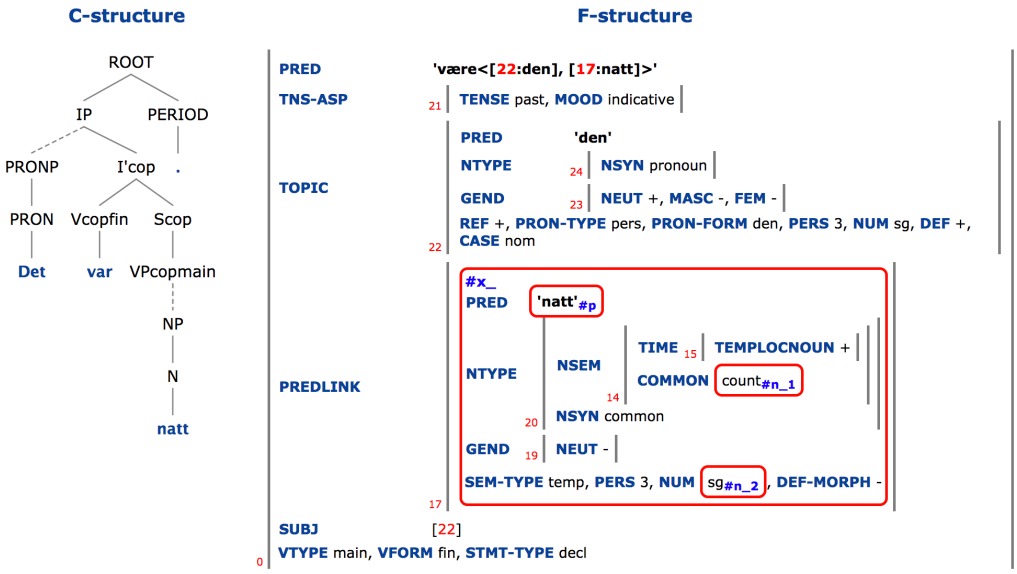


Figure 3: C- and f-structures for the sentence *Det var natt.*

4.2 Can nouns with little descriptive content occur as bare singulars?

In order to investigate whether it is true that very general nouns tend not to appear as bare singulars, we searched in the entire automatically disambiguated treebank for Norwegian Bokmål, since it is important to search through large amounts of data to detect a potentially rare phenomenon. The only restriction was that we did not search

among fragment analyses, since these will often assign incorrect bare singular analyses to nouns. The search expression in (24) was amended by adding the desired PRED form to the first conjunct: #x_ >PRED #p:'dings'. This search resulted in eight hits. One of these was in a sentence that had received an incorrect analysis of the phrase *en såpass diger dings* 'such a large thing', which is not a bare singular. One occurrence was in a headline, where noun phrases are often abbreviated with bare singulars that are not acceptable in a normal context. Three occurrences involve the use of *dings* as a euphemism for the male sex organ and two cases involved compounds that end with the element *dings*; these do not involve the very general descriptive content that is normal with this word. Finally, the example in (26) shows a kind of play on words, where *dings* is contrasted with the nonsense word *dangs*.

(26) *Testen og eksamen ble laget av en mann, sensor kjenner*
the test and exam were made by a man, examiner knows
igjen kvinneskrift, NTH er dings, kjemi er dangs osv.
again female writing NTH is gizmo, chemistry is dangs etc.
'The test and exam were made by a man, the examiner recognizes the handwriting style of a woman, NTH is gizmo, chemistry is 'dangs', etc.'

The context of this sentence makes it clear that what is being listed here are rationalizations for why there is a lack of gender equality in education. The last two clauses contrast *dings* with *dangs* in a sarcastic manner; the idea is to say that NTH (the Norwegian Institute of Technology) and chemistry are just *this and that* in a pejorative sense. This is not a normal use of the word *dings*. In conclusion, none of the occurrences of *dings* found by this search constitute legitimate examples of bare singulars.

We performed similar searches for the words *ting* 'thing' and *greie* 'thingamajig'. These words get many more hits than *dings* — 1804 for *ting* and 437 for *greie* — simply because they are more frequent words. It has not been feasible to examine all of the hits as we did above, but browsing through them we were not able to find any genuine occurrences of bare singulars. This result must be said to support the claim that bare singulars tend to be unacceptable if their descriptive content is only very general (Borthen 2003, p. 50).

4.3 Verbs that take bare singulars as objects

A second claim in Borthen (2003) is that *have*-predicates are particularly frequent with bare singular objects. In order to find out which verbs typically take bare singular objects, the search expression in (24) must be augmented to include specifications about the verb, as shown in (27). The first conjunct says that there is an f-structure #v_ (the f-structure of the verb) that has an attribute PRED with the value #p, and the second conjunct states that this verb must have the bare singular as its object (OBJ). The third conjunct specifies that the f-structure must have an attribute VFORM; this ensures that the bare singular is the object of a verb, and not of a preposition, for example. The

other conjuncts are the same as in (24) except that #x_ >PRED #p is omitted, since the purpose here is simply to list the most frequent verbs that co-occur with bare singulars, and not the nouns themselves. In order to get the most accurate results possible, we ran the search only on the manually disambiguated part of the corpus, excluding fragment analyses.

(27)
```
#v_ >PRED #p &
#v_ >OBJ #x_ &
#v_ >VFORM &
#x_ >NUM 'sg' &
#x_ >(NTYPE NSEM COMMON) 'count' &
!(#x_ >SPEC) &
!(#x_ >DEF '\+') &
!(#x_ >PRED 'pro') &
!(#x_ >(OBL PSEM) 'part')
```

The most frequent verbs according to this search are shown in Table 1, with the *have*-predicates in boldface. Note that it is the PRED value of the verb that is listed in the table. Most words have their citation form as their PRED value, but some words have special values; the verb *være* 'to be' has 'exist' as its PRED value in presentational constructions, while the PRED values *'få#øye*på'* and *'legge#merke*til'* are the predicate names of verbal idioms (see also 4.4 below).

Count	#p: value
178	**ha** 'have'
86	**få** 'get'
35	**gi** 'give'
34	**ta** 'take'
29	exist 'be-existential'
26	få#øye*på 'catch sight of'
17	legge#merke*til 'notice'
12	**bruke** 'use'
12	si 'say'
11	holde 'hold'

Table 1: The ten most frequent verbs with bare singular objects

The data in Table 1 lend some support to the claim that *have*-predicates are particularly likely to take bare singular objects, as the four most frequent verbs on the list are *have*-predicates.[3] A search for verbs with bare singular predicative NPs (i.e.

3 However, the hits include a considerable number of sentences with bare nouns that may be argued to have a mass interpretation rather than a singular count interpretation. The reason for this is most likely

bare singulars with the syntactic function PREDLINK instead of OBJ) results in 255 matches, which all involve the verb *være* 'be'. This is less than the 345 matches with *have*-predicates in Table 1. These numbers suggest that *have*-predicates constitute the class of verbs that co-occur with bare singulars most often, even if the copular verb is included among potential verbs.

The verbs in Table 1 are frequent also with corresponding objects with the indefinite article. In fact, the first five verbs listed in Table 1 are top five also if one runs a search for the most frequent verbs that take indefinite singular objects with a specifier. In other words, bare singulars are particularly likely to occur as arguments of the most frequent verbs. What is crucial is the relative frequency of verbs with bare singular objects compared to the relative frequency of verbs with other types of objects. For instance, the present search resulted in 178 hits for the verb *ha* 'have' with bare singular objects, while the corresponding search for *ha* and singular indefinite objects with a determiner resulted in 181 hits. This means that for the verb *ha*, bare singulars constitute almost fifty percent of all singular indefinite nominal objects of the verb, given that all hits are correct. This can be contrasted to other verbs, which have a much lower percentage of bare singular objects. To conclude, Table 1 lends some support to the claim in Borthen (2003) that *have*-predicates are particularly likely to take bare singular objects, but the question requires a more thorough empirical investigation and statistical analysis to be answered firmly.

Another more crucial insight of the search for verbs that take bare singular objects is that the border between the 'conventional situation type' construction and the 'profiled *have*-predicate' construction is not as clear cut as it appears to be in Borthen (2003). Some examples of sentences found with bare singulars and the verb *ha* 'have' are provided in (28)–(31), with both the verb and the bare singular in boldface.

(28) *Skal du ikke* ***ha*** ***fest*** *da?*
shall you not have party then
'Aren't you going to have a party then?'

(29) *Den fjerde mai* ***hadde*** *jeg* ***bursdag****.*
the fourth May had I birthday
'The fourth of May was my birthday.'

(30) *Den natta* ***hadde*** *ikke Kato* ***mareritt****.*
that the night had not Kato nightmare
'That night Kato didn't have a nightmare .'

that the mass–count distinction was not encoded in NorKompLeks (Nordgård 1998), the lexical resource that is the basis for the NorGram lexicon. Mass readings have been added by the annotators in the INESS project as they have been encountered during disambiguation, but there are certainly many mass nouns which have not received the proper encoding.

(31) *Jeg* ***har*** *dessverre* ***type****, men det er ikke sikkert det varer lenge.*
I have unfortunately boyfriend but it is not certain it lasts long
'I unfortunately have a boyfriend, but it's not for sure it will last long.'

Bare singulars that occur as complements of *have*-predicates, such as those in (28)–(31), may well be part of verb phrases that denote conventional situation types such as having a party, having a birthday, having a nightmare, and having a boyfriend. This suggests that the '*have*-predicate' construction and the 'conventional situation type' do not exist side-by-side as two distinct ways of generating bare singulars, as proposed in Borthen (2003).

4.4 Bare singulars in idiomatic expressions

According to Sag et al. (2002, p. 2), multiword expressions are "idiosyncratic interpretations that cross word boundaries (or spaces)". Sag et al. distinguish between two main types of expressions: lexicalized phrases and institutionalized phrases. Lexicalized multiword expressions are idiosyncratic with respect to their syntax and/or semantics, and they sometimes contain words which do not occur in isolation. Some are fixed expressions with rigid word order, while others are syntactically flexible. Institutionalized multiword expressions have normal syntactic and semantic properties, but the words that make them up co-occur with markedly high frequency. When Borthen (2003) made the claim that most bare singulars occur in multiword expressions, it was with the first multiword category in mind, i.e. expressions where the meaning and possibly also the syntax of the expression cannot be deduced from the meaning and the syntax of the individual words and the way they are put together.

Verbal idioms that involve a verb plus an object are a common type of multiword expression; well-known examples mentioned in Sag et al. (2002, p. 5) are *kick the bucket*, *shoot the breeze*, and *spill the beans*. These all involve definite objects, but bare singulars also occur, for example *give way*, *catch fire*, and *play possum*. In order to gather evidence to investigate the claim that most bare singulars occur in lexicalized multiword expressions, we searched for combinations of verbs and bare singular objects. The search expression is the same as in (27), but with the addition of #x_ >PRED #q, since we want both the verb and the head noun to appear in the frequency list. As in 4.3 we ran the search only on the manually disambiguated part of the corpus, excluding fragment analyses. Table 2 lists the most frequent verb–noun combinations.

Six of the ten most frequent hits are analyzed as verbal idioms by NorGram, as can be seen by the predicate names that incorporate the lexical items that the multiwords consist of. The other four verb–noun combinations are also multiword expressions, although NorGram analyzes them compositionally and not as verbal idioms. But in addition to the highly frequent combinations shown in the table, the search results

Count	#p: value of V	#q: value of N	translation V+N
27	ha ‘have’	rett&right ‘right’	‘be right’
26	få#øye*på ‘get eye on’	øye ‘eye’	‘catch sight of’
17	legge#merke*til ‘lay mark to’	merke ‘mark’	‘notice’
9	få#tak*i ‘get hold of’	tak ‘hold’	‘obtain’
6	få ‘get’	melding ‘message’	‘get word’
6	ha ‘have’	råd*til ‘affordance to’	‘be able to afford’
6	ta#slutt ‘take end’	slutt ‘end’	‘end’
6	holde#øye*med ‘keep eye on’	øye ‘eye’	‘keep an eye on’
6	ha ‘have’	tid*til ‘time to’	‘have time for’
6	sette#pris*på ‘set price on’	pris ‘price’	‘appreciate’

Table 2: The ten most frequent combinations of verbs and their bare singular objects

also include many verb–noun combinations that occur only once and thus are unlikely to constitute multiword expressions. In (32)–(35) are some examples of unique combinations of verbs and bare singular nouns (with both the verb and the noun in boldface).

(32) *Jeg har* ***kjøpt*** ***interrailbillett****.*
I have bought interrail ticket
‘I have bought an interrail ticket.’

(33) *Han* ***har*** ***lys*** ***stemme*** *og snakker fort.*
he has light voice and talks fast
‘He has a high-pitched voice and talks fast.’

(34) *Å pusse tennene til barna eller* ***smøre*** ***matpakke****, ga en enorm glede.*
to brush the teeth to the children or butter sandwich package, gave a enormous joy
‘Brushing the kids’ teeth or making their lunch provided enormous happiness.’

(35) *Der* ***er*** *det* ***kø*** *ved disken.*
there is it line at the counter
‘There is a line at the counter.’

To test whether *most* bare singulars occur in lexicalized multiword expressions is difficult without a manual check of all of the hits. This is beyond the scope of the

present paper. Here we can only report on our impression from browsing through the search results, and they suggest that bare singulars — indeed — occur in multiword expressions most of the time, as claimed in Borthen (2003).

4.5 Bare singulars in non-canonical sentence structures

In Borthen (2003) it is claimed that the acceptability of bare singulars is mostly unaffected by syntactic alternations. From this it follows that it should be possible to find instances of bare singulars in sentences with non-canonical sentence structure. For instance, it is expected that bare singulars can occur as topicalized and left-dislocated objects and as the postcopular element of cleft sentences. These claims can be tested through searches in NorGramBank. Since we assumed that bare singulars in these constructions would be rare, we searched in the entire corpus, with the only restriction being that we did not search among fragment analyses.

Topicalized bare singulars can be searched for by adding the constraints in (36) to the search expression in (27).

(36) #w_ >* #v_ &
#w_ >TOPIC #x_

The first conjunct in (36) says that there is an f-structure #v_ (the f-structure of the verb) that is a sub-f-structure of another f-structure #w_. The second conjunct says that this f-structure (#w_) has an attribute TOPIC with the value #x_ (the f-structure of the bare singular). The sentences in (37)–(39) provide examples of topicalized bare singulars identified through this search.

(37) *Men* ***hønsehjerne*** *kan de ha selv!*
but hen brain can they have self
'Let them be birdbrains!'

(38) *Men* ***personlig rådgiver*** *kan jeg velge selv.*
but personal adviser can I choose self
'But I can choose a personal adviser myself.'

(39) *For* ***jordkjeller*** *hadde farmor også hatt.*
for earth cellar had grandma also had
'For Grandma had also had a cellar.'

These data show that bare singulars, just as other objects, can be moved from their base position and placed in topic position.

Left dislocation differs from topicalization in that the left-dislocated constituent cooccurs with a coreferential pronoun in subject or object position in the sentence. When the left-dislocated constituent is a bare singular, the coreference relation does not exist at the token level, but at the type level. That is, the left-dislocated constituent and

the coreferential pronoun refer to the same type of thing. Left-dislocated bare singulars can be identified by simply adding #x_ >ADJUNCT-TYPE 'left-disloc' to the search expression in (24). This constraint states that the bare singular must have the attribute-value pair ADJUNCT-TYPE 'left-disloc' in its f-structure. The sentences in (40)–(42) are examples of left dislocation found by this search expression.

(40) ***Skole*** *– det er OK for noe, som å bli lærer eller*
school that is OK for something like to become teacher or
forsker.
researcher
'School – it's OK for something, like becoming a teacher or a researcher.'

(41) *Men* ***statsministerbolig****, det kan minne om stormannsgalskap?*
but prime minister residence that can remind of megalomania
'But the residence of the prime minister, isn't that reminiscent of megalomania?'

(42) ***Kniv*** *i ranselen, det kunne være livsfarlig.*
knife in the satchel that could be deadly
'A knife in the satchel, that could be deadly.'

As for clefted bare singulars, these can be identified by adding #y_ >FOCUS #x_ to the search expression for bare singulars in (24). This uniquely identifies the element in the postcopular position of cleft sentences. Some of the resulting sentences are presented in (43)–(45) below.

(43) *Det var først og fremst* ***boktyv*** *mor var.*
it was first and foremost book thief mother was
'It was above all a book thief my mother was.'

(44) *Det var* ***hund*** *hun ønsket seg.*
it was dog she wished herself
'It was a dog she was wishing for.'

(45) *En stund var han overbevist om at det var* ***maler*** *han*
a while was he convinced about that it was painter he
ville bli.
wanted become
'For a while he was convinced that it was a painter he wanted to become.'

In sum, we have found new evidence in favor of the claim that Norwegian bare singulars can take part in syntactic alternations and appear in various types of non-canonical sentence types, here exemplified by topicalization, left-dislocation and clefting. Despite the fact that we have presented only a handful of examples, the new data are more convincing than what was provided in Borthen (2003), since authentic examples (that the reader can consult) must be said to constitute more convincing evidence than invented examples whose acceptability is merely judged by the researcher.

The complete set of output sentences that resulted from the searches presented in this section do contain some undesired hits (wrong analyses), as can be expected when complicated constructions in natural language are automatically parsed and stochastically disambiguated. However, since NorGramBank allows for the manual creation of subcorpora, the output may be manually cleaned if desirable. That way one may compare the relative frequency of bare singulars in the investigated constructions with other nominals, which will add yet another level of insight.

5 Conclusion

In this paper we have illustrated how bare singulars can be searched for in NorGramBank. The study has provided some (though not full) support for the following four claims about Norwegian bare singulars put forth in Borthen (2003):

1. Nouns with very general descriptive content tend not to appear as bare singulars;
2. *Have*-predicates are the most frequent bare singular-selecting verbs;
3. Most bare singulars are part of multiword expressions;
4. Bare singulars can occur in non-canonical sentence structures; for instance they can be topicalized, left-dislocated and clefted.

More importantly, the present study has revealed some fundamental problems with the original analysis of bare singulars, due to the availability of huge amounts of authentic data. One such observation is related to the fact that the machine annotations sometimes fail to pick out bare singulars uniquely. This reveals an interesting fact: in addition to phrases that clearly have a singular count interpretation and phrases that clearly have a mass interpretation, there are many indefinites that are hard to categorize as either one. Similar observations have been made by Halmøy (2016). This questions the very premise that bare singulars constitute an interesting category on their own – which in turn means that there are probably no constructions (or grammar rules or principles) that license bare singulars specifically in Norwegian.

A second observation that points in the same direction is the fact that many examples with bare singulars fit at least two out of the four 'bare singular'-licensing constructions proposed in Borthen (2003). This is evidence against the assumption that the four alleged constructions proposed in Borthen (2003) exist side-by-side.

The data presented in this paper may be taken to indicate that bare singulars are not licensed through a set of constructions that are part of the grammar, but rather constitute a phenomenon on a par with the choice between an indefinite or a definite article. If so, the idea that determiners are obligatory for all nominal arguments in Germanic and Romance languages (see e.g. Longobardi, 1994) is threatened. Another possibility is that the secrets of bare singulars lie in the understanding of multiword expressions, ranging from fixed, non-compositional lexicalized phrases to fully compositional institutionalized ones whose status as multiword expressions relies solely on their frequency. Further studies of bare singulars – most likely based on large-scale searchable corpora – will show which of these, or other, approaches will be most successful.

Acknowledgments

We would like to thank our reviewers for their helpful comments and suggestions. We would also like to acknowledge the outstanding work that Helge Dyvik has done with NorGram and NorGramBank. Without his efforts through many years, studies such as the one reported on in this paper would not be possible.

References

Asudeh, Ash and Line Hove Mikkelsen (2000). "Incorporation in Danish: Implications for Interfaces". In: *Grammatical Interfaces in HPSG*. Ed. by Ronnie Cann, Claire Grover, and Philip Miller. CSLI Publications, pp. 1–15.

Borthen, Kaja (2003). "Norwegian bare singulars". PhD thesis. Trondheim: Norwegian University of Science and Technology.

Bresnan, Joan (2001). *Lexical-Functional Syntax*. Malden, MA: Blackwell.

Dalrymple, Mary (2001). *Lexical Functional Grammar*. Vol. 34. Syntax and Semantics. San Diego, CA: Academic Press.

De Swart, Henriette and Joost Zwarts (2009). "Less form – more meaning: Why bare singular nouns are special". In: *Lingua* 119, pp. 280–295.

Delsing, Lars-Olof (1993). "The internal structure of noun phrases in the Scandinavian languages. A comparative study". PhD thesis. University of Lund.

Dyvik, Helge (2000). "Nødvendige noder i norsk: Grunntrekk i en leksikalsk-funksjonell beskrivelse av norsk syntaks [Necessary nodes in Norwegian: Basic properties of a lexical-functional description of Norwegian syntax]". In: *Menneske, språk og felleskap*. Ed. by Øivin Andersen, Kjersti Fløttum, and Torodd Kinn. Oslo: Novus forlag, pp. 25–45.

Dyvik, Helge, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse, and Martha Thunes (2016). "NorGramBank: A 'Deep' Treebank for Norwegian". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari,

Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis. ELRA. Portorož, Slovenia, pp. 3555–3562.

Halmøy, Madeleine (2016). *The Norwegian Nominal System*. Berlin/Boston: Walter de Gruyter.

Kallulli, Dalina (1999). "The Comparative Syntax of Albanian: On the Contribution of Syntactic Types to Propositional Interpretation." PhD thesis. University of Durham.

Kiefer, Ferenc (1994). "Noun incorporation in Hungarian". In: *Acta Linguistica Hungarica* 40.1–2, pp. 149–177.

Meurer, Paul (2012). "INESS-Search: A search system for LFG (and other) treebanks". In: ed. by Miriam Butt and Tracy Holloway King. LFG Online Proceedings. Stanford, CA: CSLI Publications, pp. 404–421.

Nordgård, Torbjørn (1998). "Norwegian Computational Lexicon (NorKompLeks)". In: *Proceedings of the 11th Nordic Conference on Computational Linguistics (NoDaLiDa), Copenhagen*.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer, and Helge Dyvik (2012). "An Open Infrastructure for Advanced Treebanking". In: *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*. Ed. by Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco. Istanbul, Turkey, pp. 22–29.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger (2002). "Multiword Expressions: A Pain in the Neck for NLP". In: *Lecture Notes in Computer Science. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*. Vol. 2276. Springer, pp. 1–15.

Schmitt, Cristina and Alan Munn (1999). "Against the nominal mapping parameter: Bare nouns in Brazilian Portuguese". In: *Proceedings NELS* 29, pp. 339–353.

Stvan, Laurel Smith (1998). "The Semantics and Pragmatics of Bare Singular Noun Phrases". PhD thesis. Northwestern University.

The concept of 'translation unit' revisited

Martha Thunes

Abstract. In translation studies, the theoretical concept of 'translation unit' has traditionally been a subject of debate. This paper will discuss different views of the concept, relating it to the dichotomy between product and process-oriented translation studies. It will be argued that 'translation unit' has two readings: 'unit of analysis' in product-based studies, and 'unit of processing' in cognitive translation studies. With the exception of literary translation, translation services may now be said to fall within the domain of the language industry, which calls for considering the relevance of 'translation unit' to machine translation (MT). From a historical perspective, the concept will be related to the main issues of system design and translation quality.

1 Introduction

In theoretical studies of translation, the concept of 'translation unit' has been widely debated. While translation theory covers written as well as oral translation, or interpreting, the discussion here is limited to written translation. Some previous explications of 'translation unit' will be reviewed before various approaches to translation research are presented. In Section 2, different types of textual translation correspondences will be illustrated by samples of parallel texts. Section 3 presents the dual nature of the concept of 'translation unit', discusses the different approaches to it within product and process-oriented studies respectively, and comments on the relevance of the concept to the language industry. Finally, in Section 4 'translation unit' will be related to the field of machine translation before conclusions are drawn in section 5.

1.1 The concept

Explications of 'translation unit' provided by selected reference works may provide a point of departure for the discussion. In their *Dictionary of Translation Studies*, Shuttleworth and Cowie (1997, p. 192) define *translation unit* as "[a] term used to refer to the linguistic level at which ST [source text] is recodified in TL [target language]". Further, this is discussed in relation to Barkhudarov's definition of 'translation unit' as "the smallest unit of SL which has an equivalent in TL" (Barkhudarov 1969, cited by Shuttleworth and Cowie 1997, p. 192). In Barkhudarov's understanding, any kind

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1331.

of linguistic unit, ranging from the smallest building blocks of the language system to the level of entire texts, may occur as units of translation. This calls for a clarification: translation units are *tokens* of linguistic types, not units of the language systems.

Shuttleworth and Cowie (1997, p. 192) comment on Barkhudarov's definition by observing that "[t]he wording at a given point in ST would determine the most appropriate unit of translation, which could be expected to vary in the course of a text or even a single sentence". Thus, it is the specific translation task that determines the size and linguistic type of a translation unit. Moreover, with regard to the size of translation units, Shuttleworth and Cowie (1997, p. 192) cite Koller (1992, p. 100) who argues that the degree of structural relatedness between source and target language may influence the size of translation units. It is likely that translation between unrelated languages will involve larger units than translation between closely related languages.

In an article in *Translation. An International Encyclopedia of Translation Studies* (Kittel et al. 2004), the translation theorist Irma Sorvali states that *translation unit* "… is usually taken to denote a unit or part of the text on which the translator concentrates at one time before going on to translate the next, similar unit" (Sorvali 2004, p. 355). In this way she assumes that units of translation are processed in sequence, and one at a time. Sorvali's definition is not very specific, which is in line with her observation that it is difficult to find a definition of this concept which is "generally applicable" (Sorvali 2004, p. 355).

These introductory references are examples of various understandings of 'translation unit' offered by translation theorists. In Section 3, the concept will be explored further while relating it to different approaches to the study of translation.

1.2 Approaches to translation studies

The field of translation theory is wide and heterogeneous, and there are several possible ways of describing and classifying the various approaches to the study of translation. A standard reference in this respect is the so-called map of translation studies provided by Holmes (1988), and further articulated by Toury (1995, p. 19). According to the 'Holmes-Toury map', the discipline branches into two main subfields: pure and applied translation studies. The latter field is directed towards translation practice, whereas pure translation studies investigate the phenomenon of translation itself. This subdiscipline branches further into theoretical and descriptive translation research. Descriptive studies deal with existing translations, seeking to detect generalisations explaining the phenomenon of translation. Finally, the map presents three subfields of descriptive translation studies, oriented towards the product, process and function of translation, respectively. Over the years, the dichotomy between product-oriented and process-oriented studies has received much attention within translation research.

In short, product-oriented studies of translation are focused on topics such as characteristics of translated texts, and relations between source and target texts, whereas process-oriented studies deal with the translation activity, including the cognitive pro-

cesses behind the production of a translation. A basic difference between these two approaches follows from the fact that the product of translation is a more easily accessible object of study than the activity that takes place in the mind of the translator at work. Hence, the two approaches rely on very different research methods. In terms of methodology, a good deal of product-oriented research resembles contrastive language studies, a frequent common denominator being the use of language corpora and associated search tools. There are also product-oriented case studies which do not involve corpora. The methods of process-oriented translation research, on the other hand, are related to those of cognitive science, psycholinguistics in particular. The two approaches will be further discussed in Sections 3.2 and 3.3, respectively.

It should be noted that the division between product and process orientations could be seen as a continuum rather than as a binary distinction. Thunes (2011, pp. 18–26) presents an overview of this continuum, which shows that some theorists have described the product of translation partly by paying attention to the steps leading from source to target text, and others have described the process, but to some extent in terms of the relation between original and translation. This is in line with Chesterman (2005, p. 19) who makes the point that many translation researchers are not entirely "clear about whether the focus is on processes themselves or the results of processes".

2 Two samples of parallel texts

Before the discussion of translation studies continues, two short samples of translationally parallel texts will be presented in order to illustrate various types of textual translation correspondences. The examples are taken from two quite different domains, legislation and fiction, and demonstrate substantial text-typological differences. They are chosen as representatives of restricted and unrestricted text types, respectively.

2.1 Law text

The first example is a short piece of translationally parallel law texts: Article 91 of the *Agreement on the European Economic Area (EEA)*, and its Norwegian translation.[1]

1 Texts obtained from The Norwegian Royal Ministry of Foreign Affairs, 1992.

Article 91.	Artikkel 91.
1. The office of President of the EEA Council shall be held alternately, for a period of six months, by a member of the Council of the European Communities and a member of the Government of an EFTA State.	1. Et medlem av Rådet for Det europeiske fellesskap og et regjeringsmedlem fra en EFTA-stat skal etter tur være formann i EØS-rådet i seks måneder.
2. The EEA Council shall be convened twice a year by its President. The EEA Council shall also meet whenever circumstances so require, in accordance with its rules of procedure.	2. EØS-rådet skal innkalles av formannen to ganger i året. EØS-rådet skal også møte når omstendighetene krever det, i samsvar med forretningsordenen.

In this example there are one-to-one correspondences between translationally parallel sentences and headings. The correspondences exist at the level of main sentences, delimited by capital letters and full stops. Below the level of main sentences, the texts are no longer fully matched in terms of linguistic structure. A syntactic example may illustrate this: in paragraph 1, a passive construction in the English sentence (*The office of President … shall be held … by a member …*) corresponds to an active sentence in the Norwegian version (*Et medlem … skal … være formann …* 'a member shall be president'). However, in this sentence pair, as well as in the succeeding pairs, there is no translational mismatch between the sentences at the semantic level, and pairwise they have the same legal interpretations.

The given text sample illustrates the way in which parallel law texts are perfectly matched with respect to how the texts are divided into articles, numbers and sentences. This follows from the strict, institutionalised norms of law text writing, in particular the fact that the sequential order of the elements in a law text is of legal importance (cf. Bhatia 2010, pp. 38–39; Cao 2007, pp. 13–14; Šarčević 2007, p. 46). Hence, it is obligatory that the order of articles, paragraphs and sentences is the same in different language versions. In the case of the *EEA Agreement*, this requirement is indispensable, since the two texts have equal legal status. The Norwegian version is not, in the legal sense, a translation, but an authentic, independent law text, even if the Norwegian version has, in practice, been translated, primarily from the English text.

In parallel law texts, it appears trivial to identify translation correspondences between headings, paragraphs, and sentences. They are aligned units at the surface level of the texts, and as such they illustrate how legal translation is constrained by domain-specific text norms. Whether they qualify as units of translation cannot be decided by looking at the parallel texts alone. Correspondences between paragraphs and sentences appear as inadequate units of analysis if one aims at a deep exploration of law text translation.

From a cognitive point of view, law text translators clearly do not always work with only one sentence at a time, from top to bottom. It is necessary to consider a sequence of sentences simultaneously, and also to move back and forth in the text in order to secure consistency in the translation of recurring text elements. How this happens is not possible to detect merely by inspecting the translation result.

2.2 Fiction text

The second example is the opening of Doris Lessing's (1985a) novel *The Good Terrorist* and Kia Halling's translation (Lessing 1985b) into Norwegian:

THE house was set back from the noisy main road in what seemed to be a rubbish tip. A large house. Solid. Black tiles stood at angles along the gutter, and into a gap near the base of a fat chimney a bird flew, trailing a piece of grass several times its length.	Huset lå litt tilbaketrukket fra hovedveien, midt i noe som minnet om en søppelfylling. Et stort hus. Massivt. Svarte takstein hadde kilt seg fast i uryddige vinkler langsmed takrennene, og oppe ved skorsteinen gapte et mørkt hull; en fugl smatt inn i hullet med et strå i nebbet, strået var flere ganger lengre enn den vesle fuglekroppen.

Some comments are in order about the correspondences between sentence-level units in this piece of fiction text. Firstly, the opening sentence of the English text, *THE house was set back from the noisy main road in what seemed to be a rubbish tip*, corresponds with the opening sentence in the Norwegian translation. Then, the noun phrase *A large house* corresponds with the Norwegian noun phrase *Et stort hus*. Next, there is a correspondence between two single-word expressions, the English adjective phrase *Solid* and the Norwegian adjective phrase *Massivt*. But after this, there is a break in the pattern of one-to-one correpondences between units delimited by capital letters and full stops. The last sentence in the English text is a sequence of two conjoined independent sentences, including a non-finite adverbial subclause embedded in the second main clause: *Black tiles stood at angles along the gutter, and into a gap near the base of a fat chimney a bird flew, trailing a piece of grass several times its length.* This has been translated into a sequence of no less than four sentences in the Norwegian text, running from *Svarte takstein* onwards, and throughout the given text sample.

As in the case of the law text example, it is clearly limited to what degree the orthographic units of source and target text may serve to identify the units of processing during translation. Within these pairs of textual units, there are several instances of source–target matches as well as mismatches at various linguistic levels, especially on the level of semantics.

An example of a semantic mismatch is the correspondence between the two noun phrases *the noisy main road* and *hovedveien* 'the main road', where the information ex-

pressed by *noisy* is lost in the target text expression. Further, concerning the description of the chimney mentioned in the text, the translator has added the information that the gap close to the chimney is dark, while, on the other hand, the width of the chimney, indicated by the adjective *fat*, is not expressed in the translation. Moreover, the Norwegian verb phrase *hadde kilt seg fast* ('had got stuck') adds information not contained in the source text.

Concerning the level of syntax, several mismatches have already been mentioned in the previous discussion of sentence-level units. A further example could be the translational correspondence between the noun phrase *a gap near the base of a fat chimney* and the independent sentence *oppe ved skorsteinen gapte et mørkt hull* ('up by the chimney a dark hole was gaping'). Semantic differences involved in this example have already been commented on.

Again, the comparison of source and target text in this example can hardly shed light on what the translation units were during the translation process. For example, did the translator focus on only one of the two conjoined sentences at the end of the English text sample, or did she treat them as one unit? It is likely that she did not treat them as one unit, since they ended up as a longer sequence of four independent sentences in the translation. And when she translated the very short units *A large house* and *Solid*, did she pay attention to any neighbouring units or not?

The discussion of the two examples given of parallel texts ties in with Shuttleworth and Cowie's comments on Barkhudarov's understanding of 'translation unit', presented in Section 1.1 (Shuttleworth and Cowie 1997, p. 192). In general, a large variety of linguistic types may occur as translation units in Barkhudarov's sense: single lexical units, phrases, clausal constructions, independent sentences, paragraphs, or other types of units, large or small, depending on the particular translation task. It appears likely that this holds for the cognitive units as well as for the textual ones, and some light may be shed on this by the discussion of cognitive research on the translation process in Section 3.3.

3 'Translation unit': approaches and applications

In relation to the concept of 'translation unit', it will be shown that whether the object of study is the product or the process of translation in fact changes the content of this concept. It will be argued that within product-oriented approaches 'unit of translation' can be understood as 'unit of analysis', whereas in process-oriented studies, it primarily means 'unit of processing'.

3.1 A dual concept

The presentations of product-oriented and process-oriented studies will reveal that the concept of 'unit of translation' has a dual nature, and this view is compatible with a definition given by two researchers who work within the avenue of process-oriented translation studies, Amparo Hurtado Albir and Fabio Alves. In *The Routledge*

Companion to Translation Studies (edited by Jeremy Munday 2009), Hurtado Albir and Alves describe 'unit of translation' both as a "(bi)textual unit", and as a cognitive unit. From the perspective of the translation process, they understand the concept as a "[c]ommunicative and cognitive unity [sic] employed by a translator/interpreter in the performance of a translation task" (Munday 2009, p. 238). In other words, this is a 'unit of processing'. From a textual perspective, they claim that the concept "is embedded in a complex relationship with all the other units in a given text" (ibid.). The phrase "other units" refers to micro-textual units, as well as units of a macro-textual character, and units with special text-structuring functions. This calls for the 'unit of analysis' reading, as the identification of such textual units presumes some kind of linguistic analysis.

3.2 Product-oriented approaches to 'translation unit'

One definition of 'unit of translation' in the context of product-oriented studies is provided by Malmkjær 1998 (p. 286): "Considered from a product-oriented point of view, the unit of translation is the target-text unit that can be mapped onto a source-text unit". Prominent topics within the product-oriented approaches have been, e.g., characteristic features of translated texts, the relationship between source and target texts, and comparisons of different translations of the same originals, whether into one or more languages. Such studies have in common that the researcher's observation applies to intersubjectively available objects, i.e. the translated text in comparison to the source text. These are entities that have been produced before the observation takes place.

An example of a strictly product-oriented approach is found in Thunes (2011), which is a study of linguistic correspondences in English-Norwegian parallel texts of two types, law and fiction. The empirical analysis is based on the assumption that it is possible to compute, or construct without human intervention, a target language expression by using information about source and target language systems, and about how the two language systems are interrelated. This is an assumption that lies at the bottom of linguistic approaches to machine translation.[2]

The main research questions in Thunes (2011) are, first, to what extent it is possible to compute the translations founds in the selected parallel texts, and, second, whether the chosen text types differ with respect to the first question. In the empirical investigation, the notion of 'translation unit' is used purely in the sense of 'unit of analysis'. The discussion of empirical data accentuates the product-oriented approach of Thunes (2011), as the analysed units are generally referred to as *string pairs*, or *translational correspondences*, and not as *translation units*.

In the study, the finite clause is chosen as the primary unit of analysis, and the motivation for this is twofold. Firstly, finite clauses can be identified by fairly simple linguistic criteria, which makes it easy to detect relevant patterns in the analysed

2 Cf. Section 4.2.

parallel texts. Secondly, the finite clause is chosen because it reflects the approach of rule-based machine translation, which normally works at sentence level. For methodological purposes, phrases with embedded clauses are also applied as units of analysis.

Elgemark (2017) is another example of the product-oriented approach to translation. Her investigation is a corpus-based, contrastive study of English–Swedish, aimed at exploring a phenomenon related to the information structure of sentences, i.e. *N-Rhemes*, defined by Elgemark (2017, p. 1) as "the last constituent that has a function in the clause". Normally, N-Rhemes contain information that is of relatively high importance in utterances. Elgemark's study aims at describing properties of N-Rhemes individually in the two languages, as well as examining correspondences and non-correspondences between English and Swedish N-Rhemes.

As a unit of analysis, she applies *T-units*, which are, basically, main clauses including any embedded dependent clauses, and in the empirical investigation, the N-Rheme is identified in each T-unit. Elgemark's analysis of corpus data has detected word order and information structure differences between English and Swedish N-Rhemes, mirroring differences between the two language systems.

The translation theorist Gideon Toury applies an understanding of translation unit as a unit of comparative analysis, or in Toury's own terminology, "the coupled pair". Toury's contribution to descriptive translation studies can be placed among the approaches that are not purely product-oriented, but show elements of process orientation. He describes his study as "an attempt to gradually reconstruct both translation decisions and the constraints under which they were made" (Toury 1995, p. 88), and this is his motivation for identifying units of comparative analysis.

He defines 'coupled pairs' as correspondences between specific translation problems in the source text (i.e. tasks to be solved), and their solutions in the target text (1995, p. 77). Also, he emphasises that in coupled pairs, source problems and target solutions "should be conceived of as determining each other in a mutual way" (1995, p. 77).

The perhaps most noticeable aspects of Toury's understanding of 'translation unit' are, firstly, that it involves a *pair* of linguistic segments, and, secondly, that the two units of analysis mutually determine each other. It is a challenging question to what extent such pairings can reveal the decisions made by translators, because the actual correspondences that we find in translationally parallel texts are created by an interplay between the translational relationship between source and target language systems, on the one hand, and, on the other hand, a range of factors which are specific to individual translation tasks.

3.3 Process-oriented approaches to 'translation unit'

In process-oriented studies of translation, the object of study is the translator's activity. From the cognitive perspective, Hurtado Albir and Alves state that "… the translation unit is considered as a comprehension unit and as a processing unit, i.e. as a dynamic segment of the ST [source text], independent of specific size or form, to which, at a

given moment, the translator's focus of attention is directed …" (cf. Munday 2009, p. 238). This description implies that the processing of a translation unit relies on its comprehension. Thus, if the translator's understanding of the source text can be regarded as a kind of analysis, it follows that the analysis aspect is present also in this view of 'translation unit'. However, the dynamic character of the processing unit is a more prominent aspect, and the unit is dynamic in the sense that its length and linguistic type may vary as the translator is working.

Concerning the translation process, research has shown that it cannot be assumed that there is a fixed set of steps that are carried out in any act of translation. This was reported already by Krings' (1986) dissertation *Was in den Köpfen von Übersetzern vorgeht*, which was the first extensive, published study of translation activity using the method known as Think-Aloud-Protocols (TAP). In TAP studies the translator is typically asked to report, unselectively, everything that goes through her/his mind when performing the translation task, i.e., literally, to think aloud, while the reporting is audio or video recorded. Other actions, such as note-making and consulting reference works, are also documented.

TAP studies have shown that the ways in which translation processes may run are influenced by numerous factors determined by the skills of the translator, by the translation situation, and by the type of translation task, to mention some. In the light of this, it is to be expected that the unit of translation, or unit of processing, also varies greatly during translation activity. This was documented fairly early by TAP studies. Malmkjær (1998, p. 286) observes that "… Lörscher (1991, 1993) shows that the unit of translation used by language learners tends to be the single word, while experienced translators tend to isolate and translate units of meaning, normally realized in phrases, clauses or sentences".

In this context, it is a relevant question whether translators are aware of the actual units of processing. Sorvali (1994) tried to investigate, among other things, how translators perceive the unit of translation, whether the unit really exists during translation, and whether translators make use of it. She concluded that "… translators regard the unit of translation as a self-evident fact to which they rarely give any thought during the actual translation process but which nevertheless exists as a functional unit" (Sorvali 2004, p. 358).

This raises another question: what is the significance and status of the concept of 'translation unit'? Sorvali (2004) notes that there may be different answers to this among translators and translation researchers. Among translators, 'translation unit' is associated with practical work, but it may not be of great significance: according to Sorvali (2004, p. 355), "[t]he unit of translation may … be no more than an insignificant intermediate stage in the process as far as the translator is concerned". She further argues that although the unit of translation is a more abstract concept among translation theorists than among translators, it is still more significant to the former than to the

latter, because "[r]esearchers may use such a unit as a tool for providing a theoretical description of the translation process". Also, she explains that the types of linguistic units used for such descriptive purposes may be of different kinds than the units that are actually processed by the translator at work (2004, p. 355).

Sorvali (2004, p. 358) presents some conclusions regarding the translation unit in the context of the translation process. Firstly, because the unit of processing is so variable, or instable, and because it is so highly dependent on the translation task and situation, it is difficult to give the concept a general definition. Secondly, it cannot necessarily be deduced from the product of an act of translation what kinds of units that have been used during the process. Thirdly, the quality of the product indicates whether the translator has selected units of a size that is sufficient in order to create a successful target text. The argument is that if the units of processing are insufficient, the translator may not be able to choose optimal target expressions.

The latter point is compatible with an observation made by Malmkjær (1998, p. 286), regarding translation quality: "The typical finding is that target texts in which the units are larger appear more acceptable than those in which the units are smaller." Further, Malmkjær (1998, p. 286) concludes that the clause is the primary unit of translation: "In general, the clause seems a sensible structure to aim for as translation unit, because it tends to be at clause level that language represents events […] In addition, the clause is a manageable unit of attentional focus […]".

This reference to attentional focus leads over to the strictly process-oriented study of Carl and Kay (2011), which is worth looking into in some detail as it has provided ground-breaking insight into translation. Their contribution is an exploration of the cognitive notion of 'translation unit', based on the observation that "… there is a confusion in the usage of the term *translation unit* …" because some researchers apply it to "… basic segments of activities in the *translation process*, whereas others think of the segments more statically as properties observable in the *translation product* …" (Carl and Kay 2011, p. 953). Their solution to this terminological problem is to reserve *translation unit* for "units of cognitive activity", defined as "the translator's focus of attention". These units can be explored in data on the translation process. Further, they use the term *alignment unit* to refer to translational correspondences observable in the product of translation (ibid.).

Carl and Kay (2011, p. 960) assume three phases in the translation process: skimming the source text (ST), drafting the target text (TT), and revising the translation. In their study, the reading and writing activities of translators at work are documented by tracking their eyeball movements during skimming, and by tracking their typing of the draft translation. The Translog software (Jakobsen 1999) is used for acquiring such *user activity data*. To achieve a controlled experiment, the subjects used no dictionaries or other common translation tools.

In order to analyse these data as empirical facts about the cognitive process of translation, Carl and Kay (2011, p. 954) "… rely on the "eye-mind assumption" (Just and Carpenter 1980), which hypothesizes that "there is no appreciable lag between what is being fixated and what is being processed" (Just and Carpenter 1980, p. 331). On the basis of this assumption, Carl and Kay (2011, pp. 955–956) identify *fixation units* during skimming by tracking and analysing the translator's eyeball movements when reading the source text.

With respect to the translator's typing during drafting, Carl and Kay (2011, p. 954) assume "… it is likely that the translator's focus of attention is close to what s/he writes, and that units of text production coincide to some extent with the entities of the translator's cognitive processes". This assumption allows for identifying *production unit*s by logging the translator's typing during drafting (2011, p. 955). Notably, their term *production unit* signifies the dynamic, cognitive concept; *product unit*, which would be synonymous with the static concept 'alignment unit' does not occur in their work.

Thus, Carl and Kay (2011) decompose the translation unit into two different kinds of processing units, the fixation unit involved in ST understanding, and the production unit of TT writing. They describe the physical realisations of translation units as three-component structures (2011, p. 954): a translation unit consists of (i) an act of writing which creates a production unit within a certain time span, (ii) an act of reading which tells the translator how to translate the fixation unit which is read, and (iii) "the ST segment(s) of which the produced TT is a translation". Surprisingly, the third component refers to alignment units, which, we have seen, are excluded from Carl and Kay's concept of 'translation unit' (2011, p. 953).

The method applied in order to identify translation units in the recorded user activity data relies on mappings between different types of information (Carl and Kay 2011, p. 957): eyeball movements are mapped onto ST fixation units; typing actions onto TT production units. Then, by relating these mappings to source–target alignment information, it is possible to identify correspondences between individual keystrokes and specific ST units. Thus, Carl and Kay (2011) have provided an exact, empirical method for describing how processing units are realised during the act of translation.

An important aspect of this method is the tuning of the production unit segmentation threshold (Carl and Kay 2011, pp. 966–969). This amounts to setting the minimum length of a typing pause that may identify a boundary between two production units. Selecting a too low threshold value will yield segments of an arbitrary character, i.e. segments which are neither cognitively nor linguistically plausible as they do not form complete units of meaning. Carl and Kay (2011, p. 969) conclude that "[t]he likelihood of [production units] to be consistent with linguistic entities … is maximal for typing pauses of one second or more". In relation to Malmkjær's (1998, p. 286) assumption that the clause is the primary unit of translation, it is interesting that Carl and Kay's (2011) findings show that production units will not necessarily represent complete

units of meaning, and need not conform with alignment units (2011, p. 956). On the other hand, this supports the view of Sorvali (2004, p. 355), commented on above, that units of analysis may be of different kinds than units of processing.

A great asset of Carl and Kay's (2011) method of analysing fixation and production units is that it provides intersubjectively available data on the processes that take place in the minds of individual translators. However, fixation as well as production units are not unique to the translation process; they occur generally in reading and writing activities, such as text copying.[3] The Translog system used by Carl and Kay (2011) to acquire translation process data is designed for tracking any kind of computer-based reading and writing activity. In their study, it is by linking the recorded fixation and production units with alignment units that it becomes possible to identify the units of the translation process. This illustrates the point made by Thunes (2011, p. 66) that "… in the case of translation it is the product and its relation to the original text which gives the process its identity". Still, this observation does not reduce the importance of Carl and Kay's (2011) contribution to knowledge about the translation process, and their study is a very good answer to a call for conceptual clarity made by Chesterman (2005, pp. 17–22), where he argues that many concepts applied in translation research need to be sharpened in order to achieve terminological stringency across the field.

3.4 Relevance to the language industry

Moving out of the domain of translation research, it may be noted that translation has become an everyday tool, an application taken for granted by everyone with Internet access. In a historical perspective, translation started as something that was carried out by relatively few people of high learning, and applied only to texts of special importance. Towards the modern age, translation gradually emerged as a profession to meet a growing market, and during the 20th century, globalisation created an enormous demand for non-literary translation in multinational domains of industry, trade, legislation, politics, and science.

Thus, translation has become part of the language industry. Machine translation in the shape of Google Translate has become as widespread as the Internet, and has achieved a position where it can even form the layperson's idea of what translation is. In spite of its usefulness, this application cannot, however, remove the need for translation work carried out by professional, human translators. On the other hand, translators have become dependent on language technology in order to meet demands of efficiency. Bilingual dictionaries, terminology bases, and translation memories are among the required parts of a translator's work station. Moreover, given the availability of useful MT systems, there has been an important change in the way professional translation is carried out. With the exception of literary translation, it has become a normal approach to post-edit machine-translated text rather than to work from scratch.

3 Cf. the study by Carl and Dragsted (2012) where translation is compared with text copying.

As discussed in Section 3.3, the concept of 'translation unit' is hardly of any relevance to the translator at work. For the translator, as well as for the client, what matters is the quality of the output, not the size or linguistic type of individual processing, or production units. The product-oriented notion of 'translation unit' is, however, applicable when distinguishing between different types of translation tools. In bilingual dictionaries and terminology bases, the lexical unit is the primary unit of translation, whereas in translation memories, which are databases of previously translated texts aligned with their source texts, units of translation are headings, paragraphs, sentences, clauses, phrases, list items, and others.

4 'Translation unit' in machine translation

The product-based concept of 'translation unit' may shed some light on certain issues of the domain of automatic translation. Jurafsky and Martin (2009, p. 898) divide the field into classic and modern machine translation, and this opposition reflects the important distinction between *rule-based MT* (RBMT) and *statistical MT* (SMT). In recent years, a newer approach has evolved, too, *neural machine translation* (NMT). SMT and NMT may be described as non-linguistic methods, in contrast to the linguistic approach of RBMT. In Sections 4.1 and 4.2, the concept of 'translation unit' will be related to different paradigms within MT research and development.

4.1 Relevance to non-linguistic approaches to MT

In RBMT, the translation procedure relies on information about source and target language and their interrelations, whereas in SMT, translations are computed on the basis of statistical, or probabilistic, information about recurrent translational patterns in large bodies of parallel texts, or parallel corpora. Thus, the corpus provides what is regarded as the training data for the system; these data provide the probabilistic information which is the basis for generating the translations output to the end user. SMT techniques work without using any information about source and target languages, and this is why they are normally described as a non-linguistic approach to MT. An important reason for the success of modern, statistical MT is that it does not suffer from the problem of lexical coverage, which has been among the heaviest obstacles to the development of linguistic MT systems. While RBMT systems cannot translate words which are not included in their lexical databases, SMT systems do not need lexicon components. However, lexical coverage in a different sense is a challenge also for SMT: if a word does not occur among the training data, then an SMT system cannot compute a target language match for it.

As the concept of 'translation unit' is primarily a linguistic notion in translation theory, it appears to be of limited relevance to the field of statistical MT. One aspect could be commented on: SMT works by using probabilistic information about translational correspondences between word sequences, or N-grams, in the training corpus. An N-gram is a sequence of N words where N is a low number. The value of N may

vary between translation systems, and will have a maximum value in each system. If N is 1, 2 or 3, the word sequences are uni-grams, bi-grams or tri-grams, i.e. strings of respectively 1, 2 or 3 word forms in running text. Possibly, this notion of N-gram could be given the status of 'translation unit' in SMT, but as no linguistic analysis is involved, it is not a unit of analysis in the translation-theoretic sense. N-grams could, however, be regarded as units of processing, i.e., units of processing defined by the algorithms implemented in statistical translation systems.

Neural machine translation (NMT) draws on neural network technology, in which computational systems are modelled on biological neural networks. Like SMT, NMT relies on probabilistic techniques. NMT models are trained on representations of the entire source and target sentences, rather than on N-grams. Words are still important as units in the source and target texts, but "[c]onnections between source and target words, phrases and sentences are learnt only implicitly as mappings between their continuous representations" (Kalchbrenner and Blunsom 2013, p. 1701). Hence, in NMT, it is hard to see any applicability of the translation-theoretic notion of 'translation unit'.

4.2 Relevance to linguistic MT

It is more relevant to apply the notion of 'translation unit' to the linguistic approach of rule-based machine translation. In RBMT systems, some notion of a translational unit is applied as an analytical concept in the construction of the system. During the procedure of computing the target text, this unit becomes a unit of processing from the point of view of the algorithm used by the system. Thus, within linguistic MT, units of translation are units of analysis as well as units of processing. This is a parallel to the dual nature of the translation-theoretic concept.

From a historical perspective, there has been some variation concerning the linguistic types of translation units applied in rule-based machine translation. First-generation systems operated on word-level translation units. These systems were designed using the so-called *direct translation strategy*, which can be described as mapping the words in the input text directly onto words in the target language. Direct MT systems are commonly characterised as implementations of bilingual dictionaries with certain syntactic reordering rules for accommodating structural differences between source and target language.

Early work on machine translation grew out of information theory in the 1950s and 1960s, in a period when information scientists often held the view that translation was in essence a task of decoding the source text and recoding it in the symbols of the target language. During the same period, many translation theorists held quite similar views of what translation was about. This point is e.g. made by Koller (1992, pp. 89–92) in a discussion of early models in translation theory, and Sorvali (2004, p. 355) claims that in early translation research, the word was often regarded as the unit of translation, which she ascribes to the influence of the structuralist tradition in language studies. Clearly,

in that era there were strong parallels between machine translation and translation theory with respect to the conception of translation.

Direct MT systems failed to deliver high-quality output, and researchers concluded that it was necessary to develop systems that were able to do a more thorough linguistic analysis of the source text. The field then saw the emergence of various types of second-generation systems, which came to be described as *indirect MT systems* to separate them from the direct systems of the first generation. The development of indirect translation systems involved highly sophisticated computational linguistic engineering. Although there were substantial differences in system architecture among second-generation systems, a common denominator between them is that the sentence is normally the primary unit of translation.

The main characteristic of the indirect approach is that the first step in the translation procedure is an analysis stage which produces a formal, system-specific representation of the input sentence. In such representations, the meaning and structure of the source language expression are made more explicit than in the source expression itself. Thus, in some second-generation systems, the representation of the input sentence contains sufficient information for generating a target sentence. In other systems, the target sentence will be generated after the source text representation has been modified in accord with information about the target language system and how it is related to the source language.

There are some quite understandable reasons why the sentence is the basic unit in second-generation systems. Firstly, if a translation system is to be useful, it must at least be able to handle linguistic units at sentence level. Direct MT systems did not do this, and the quality of the output that they produced was in general too poor. Secondly, a translation system must be able to deal with sentences because the sentence can be seen as the maximal domain of grammatical analysis. That is, the sentence is the largest type of linguistic unit whose construction is governed by syntactic rules. The latter point echoes the translation-theoretic view of the clause as the typical unit of translation, which was presented in Section 3.3 with reference to Malmkjær's (1998, p. 286) comments: firstly, if translators work on units of an insufficient size, translation quality tends to suffer, and, secondly, the clause is a sensible unit because it is manageable, and functional since in natural languages events are normally represented at clause level.

On the other hand, automatic translation becomes non-robust if a system operating at sentence level does not return a translation whenever the input cannot be recognised as a syntactically complete source language sentence. Failed analyses may occur either because the ST sentence contains a grammatical structure that is not covered by the SL rule component of the system, or because the input is not a complete sentence, which frequently occurs in running text. Within rule-based MT, there are systems that process phrase-level units in cases where the input is not a sentence, but a smaller unit,

or in cases where the system is unable to analyse the input sentence completely, but is able to recognise subparts of the input as independent syntactic units. An example is LOGON (Oepen et al. 2004, 2007), a Norwegian-to-English translation system which was developed by a Norwegian research group in cooperation with international partners. Primarily, LOGON processes sentences, but the system is also designed to handle noun phrases and preposition phrases. In relation to the issue of robustness, this is a clear advantage and illustrates the fruitfulness in MT of using not only sentences as processing units, but also linguistic structures at phrase level.

5 Summary and conclusions

There are two distinct readings of the translation-theoretic concept of 'translation unit', and these are correlated with the fundamental differences between product-oriented and process-oriented approaches to translation research. Within product-oriented studies of source texts paired with their translations, the concept can be understood as 'unit of analysis'. It is a bi-textual linguistic unit, an alignment unit, and plays an important role in the methods of corpus-based contrastive language studies. Within process-oriented studies of translation activity, 'translation unit' can be read as 'unit of processing', which can further be explicated as a cognitive unit of attentional focus.

The concept of 'translation unit' is mainly of importance to translation researchers, less so for translators. In relation to the language industry, the product-oriented reading is of some relevance in relation to tools like bilingual term bases and translation memories. When it comes to machine translation, the concept is of minor relevance in SMT, and even less in NMT. However, the rule-based approach of linguistic MT can be related to both readings of 'translation unit'. In the design of an RBMT system, the 'unit of analysis' reading applies to the types of source text units that can be identified by the system, and the 'unit of processing' reading applies when the translation algorithm operates on the analysed input in order to generate a target text.

Over the years, translation theorists have argued that translation units are highly variable with respect to size and linguistic type. This observation holds for both readings of 'translation unit' across the different fields that have been discussed.

Recent research on the cognitive activities of translators has contributed to a sharpening of the distinction between the two understandings of 'translation unit'. However, it has been shown in this article that it is difficult to identify units in the translation process without relating them to textual alignment units, or product data. The reason is that the units of cognitive activity that can be observed while translators are working can be associated with reading and writing activities in general. In order to identify the actual units of the translation process, it is necessary to link process data with data collected from the translation product.

6 Acknowledgements

First, credit is due to Anna Sågvall Hein and Magnus Merkel who once challenged me to discuss the notion of 'translation unit' in a trial lecture. Second, I thank Ingrid Simonnæs, who made me bring this work further by inviting me to present it in 2012 at the annual meeting of the Association of Government-Authorized Translators in Norway. Next, two anonymous reviewers are acknowledged for valuable and motivating comments, and I am indebted to Koenraad De Smedt for important feedback, as well as swift assistance, without which this paper would not have surfaced. Moreover, I am grateful to Sandra Halverson for excellent mentoring. It has been a privilege to receive her stimulating expert advice during the writing process. Finally, Helge Dyvik deserves warm thanks for highly insightful comments on early versions of this product. Unwittingly, during the later stages, he has also provided fruitful input through interesting discussions, a typical example of his keen and generous engagement in what fellow linguists are working on. Further, I am grateful to Helge for introducing me, many years ago, to rule-based machine translation, a fascinating field dealing with linguistic delicacies such as the syntax-semantics interface in a cross-linguistic perspective. I also had the chance to work with Helge's experimental MT system PONS, and through his inspiring teaching I learnt how gratifying it is to study language from a contrastive point of view.

References

Baker, Mona, ed. (1998). *Routledge Encyclopedia of Translation Studies*. London and New York: Routledge.

Barkhudarov, Leonid (1969). "Urovni yazykovoy iyerarkhii i perevod [Levels of language hierarchy and translation]". In: *Tetradi perevodchika [The Translator's Notebooks]* 6, pp. 3–12.

Bhatia, Vijay K. (2010). "Specification in legislative writing: accessibility, transparency, power and control". In: *The Routledge Handbook of Forensic Linguistics*. Ed. by Malcolm Coulthard and Alison Johnson. London and New York: Routledge, pp. 37–50.

Cao, Deborah (2007). *Translating Law*. Topics in Translation 33. Clevedon, Buffalo, and Toronto: Multilingual Matters Ltd.

Carl, Michael and Barbara Dragsted (2012). "Inside the monitor model: Processes of default and challenged translation production". In: *Translation: Computation, Corpora, Cognition (Special Issue on the Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation)* 2.1, pp. 127–145.

Carl, Michael and Martin Kay (2011). "Gazing and typing activities during translation: A comparative study of translation units of professional and student translators". In: *Meta: Journal des traducteurs / Meta: Translators' Journal* 56.4, pp. 952–975.

Chesterman, Andrew (2005). "Problems with strategies". In: *New trends in Translation Studies: In honour of Kinga Klaudy*. Ed. by Kristina Károly and Ágosta Fóris. Budapest: Kiadó, pp. 17–28.

Elgemark, Anna (2017). *To the Very End. A contrastive study of N-Rhemes in English and Swedish translations*. PhD thesis. University of Gothenburg.

Gambier, Yves and Jorma Tommola, eds. (1993). *Translation and Knowledge: Proceedings of the Fourth Scandinavian Symposium on Translation Theory, June 4–6, 1992*. Turku: Centre for Translation and Interpreting.

Hansen, Gyde, ed. (1999). *Probing the process in translation: Methods and results*. Vol. 24. Copenhagen Studies in Language. Fredriksberg: Samfundslitteratur.

Holmes, James S. (1972/1988). "The Name and Nature of Translation Studies". In: *Translated! Papers on Literary Translation and Translation Studies*. Ed. by James S. Holmes. Amsterdam: Rodopi, pp. 67–80.

Jakobsen, Arnt Lykke (1999). "Logging target text production with Translog". In: *Probing the process in translation: Methods and Results*. Ed. by Gyde Hansen. Vol. 24. Copenhagen Studies in Language. Fredriksberg: Samfundslitteratur, pp. 9–20.

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. Upper Saddle River, New Jersey: Pearson Education / Prentice Hall.

Just, Marcel Adam and Patricia A. Carpenter (1980). "A theory of reading: from eye fixations to comprehension". In: *Psychological Review* 87.4, pp. 329–354.

Kalchbrenner, Nal and Phil Blunsom (2013). "Recurrent continuous translation models". In: *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 1700–1709.

Kittel, Harald, Armin Paul Frank, Norbert Greiner, Theo Hermans, Werner Koller, José Lambert, and Fritz Paul, eds. (2004). *Übersetzung — Translation — Traduction. Ein internationales Handbuch zur Übersetzungsforschung / An International Encyclopedia of Translation Studies / Encyclopédie internationale de la recherche sur la traduction*. Vol. 1. Berlin / New York: Walter de Gruyter.

Koller, Werner (1992). *Einführung in die Übersetzungswissenschaft*. Heidelberg: Quelle & Meyer.

Krings, Hans P. (1986). *Was in den Köpfen von Übersetzern vorgeht*. Tübingen: Günter Narr.

Lessing, Doris (1985a). *Den gode terroristen (The Good Terrorist)*. Trans. by Kia Halling. Oslo: Gyldendal.

– (1985b). *The Good Terrorist*. London: Jonathan Cape.

Lörscher, Wolfgang (1991). *Translation Performance, Translation Process and Translation Strategies: A Psycholinguistic Investigation*. Tübingen: Günter Narr.

– (1993). "Translation Process Analysis". In: *Translation and Knowledge: Proceedings of the Fourth Scandinavian Symposium on Translation Theory, June 4–6, 1992*. Ed. by

Yves Gambier and Jorma Tommola. Turku: Centre for Translation and Interpreting, pp. 195–211.

Malmkjær, Kirsten (1998). "Unit of translation". In: *Routledge Encyclopedia of Translation Studies*. Ed. by Mona Baker. London and New York: Routledge, pp. 286–287.

Munday, Jeremy, ed. (2009). *The Routledge Companion to Translation Studies*. London and New York: Routledge.

Oepen, Stephan, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, and Victoria Rosén (2004). "Som å kapp-ete med trollet? Towards MRS-Based Norwegian–English Machine Translation". In: *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Baltimore, MD.

Šarčević, Susan (2007). "Making multilingualism work in the enlarged European Union". In: *Language and the Law: International Outlooks*. Ed. by Krzysztof Kredens and Stanisław Goźdź-Roszkowski. Łódź Studies in Language 16. Frankfurt am Main: Peter Lang, pp. 34–56.

Shuttleworth, Mark and Moira Cowie (2007). *Dictionary of Translation Studies*. Manchester, UK and Kinderhook (NY), USA: St. Jerome Publishing.

Sorvali, Irma (1994). *Översättare och översättningsprocess*. Institutionen för nordiska språk vid Uleåborgs universitet.

– (2004). "The problem of the unit of translation". In: *Übersetzung — Translation — Traduction. Ein internationales Handbuch zur Übersetzungsforschung / An International Encyclopedia of Translation Studies / Encyclopédie internationale de la recherche sur la traduction*. Ed. by Harald Kittel, Armin Paul Frank, Norbert Greiner, Theo Hermans, Werner Koller, José Lambert, and Fritz Paul. Vol. 1. Berlin / New York: Walter de Gruyter, pp. 354–362.

Thunes, Martha (2011). *Complexity in Translation. An English-Norwegian Study of Two Text Types*. PhD thesis. University of Bergen.

Toury, Gideon (1995). *Descriptive Translation Studies and Beyond*. Benjamins Translation Library 4. Amsterdam and Philadelphia: John Benjamins.

Subject properties in presentational sentences in Icelandic and Swedish

Annie Zaenen, Elisabet Engdahl & Joan Maling

Abstract. We review the various non-canonical positions in which the thematically highest argument can occur in Icelandic and in Swedish. We show that NPs in these non-canonical positions have subject properties in both languages. We summarize the restrictions that we are aware of holding on the various positions and discuss whether they are configurational or thematic/semantic in nature.

1 Introduction

Scandinavian languages are considered to be strongly configurational, meaning that grammatical functions are identified with phrase structure positions. More specifically, in matrix clauses the subject appears either immediately before the tensed verb or immediate following it. We will call these positions *canonical subject positions*. Although these positions are the most common positions for subjects, it has, of course, been observed that NPs with the same thematic relation to the verb can occur in other positions; this is especially true of the indefinite NPs that occur in presentational constructions. Following e.g. Beaver et al. (2005), we will call these *pivots*. Discussions of pivots tend to center on the nature of the definiteness constraint. What has been less studied is whether pivots have syntactic subject properties or whether they show more object-like behavior. In traditional grammar, subjects are defined either by case marking and agreement properties or by positional properties. Under a positional definition of subject, pivots are obviously not subjects. Keenan (1976), however, introduced a distinction between coding properties, behavioral properties and semantic properties which allows for a more nuanced analysis. Older linguistic descriptions focussed on coding properties, but behavioral properties are those that in current linguistic theories are more often seen as being properly syntactic.

In this paper we investigate the degree to which these syntactic properties of pivots are similar to those of canonical subjects in two Scandinavian languages: Icelandic, an insular Scandinavian language, and Swedish, a mainland Scandinavian language. In the first part of the paper we argue that pivots in both languages, even those internal

The very model of a modern linguist. Edited by Victoria Rosén and Koenraad De Smedt. BeLLS Vol. 8 (2017), DOI 10.15845/bells.v8i1.1335.

to the VP, have syntactic subject properties. In the second part of the paper we show that there are some thematic constraints on these pivots that don't apply to NPs in canonical subject positions. We conclude with a discussion of how the properties we have found might be parcelled out among notions of subject and topic. Some of our findings go against previous research on mainland Scandinavian languages. For Norwegian, it has been claimed that pivots are objects (Askedal 1986; Lødrup 1999; Sveen 1996) and Mikkelsen (2002) makes the same claim for Danish. This has been questioned for Swedish by Börjars and Vincent (2005) and we elaborate here on their findings.

2 Syntactic subject properties of pivots

2.1 Icelandic

Icelandic is often presented as a configurational language par excellence because it can be shown that non-nominative NPs in the canonical subject positions do pass syntactic subjecthood tests, whereas nominative co-arguments of these NPs that are realized outside of these positions do not have these subject properties (Zaenen et al. 1985; Sigurðsson 2004). In their analysis Zaenen et al. (1985) follow Keenan (1976) in making a distinction between coding and behavioral properties of potential subjects. Coding properties are position, case marking and verb agreement. Behavioral properties are inter alia reflexivization, control and raising. Zaenen et al. (1985) took the behavioral properties as the most interesting from a syntactic point of view, so they called the NP which displayed these properties the subject. They established that in Icelandic these properties did not depend on case marking for derived subjects (more specifically, subjects in passive constructions). Their demonstration was spelled out more extensively for non-derived, basic subjects in active constructions by Sigurðsson (2004). However, Zaenen et al. (1985) as well as Sigurðsson (2004) limited their discussion to NPs in canonical subject position. They showed that these NPs have, over and above their positional characteristics, behavioral properties (control, obligatory reflexivization, raising to object (AcI) and subject ellipsis) that distinguish them from other nominal elements in the sentence, but they did not investigate whether these properties also apply to what we here call pivots. This is the question we address in this paper. Do pivots display the behavioral subject properties listed above or not? We investigate this for Icelandic as we think that the question has never been raised with respect to that language and we compare the results with results for a mainland Scandinavian language. In this article we only investigate Swedish and leave the situation in Danish and Norwegian for future research.

Like many other languages, Icelandic has a presentational construction in which an indefinite NP occurs to the right of the canonical subject positions. Icelandic even has an exceptionally rich variety of positions in which such NPs can occur with different constraints associated with each position. We summarize here the account given in

Thráinsson (2007, p. 314), who gives the following examples illustrating the various positions.

(1) *Það hafði* ***einhver köttur*** *alltaf verið í eldhúsinu.*
EXPL had some-NOM cat-NOM always been in kitchen-the
'There had always been some cat in the kitchen.'

(2) *Það hafði alltaf* ***einhver köttur*** *verið í eldhúsinu.*
EXPL had always some-NOM cat-NOM been in kitchen-the
'There had always been some cat in the kitchen.'

(3) *Það hafði alltaf verið* ***einhver köttur*** *í eldhúsinu.*
EXPL had always been some-NOM cat-NOM in kitchen-the
'There had always been some cat in the kitchen.'

We also follow Thráinsson (2007) in labeling the pivots in these positions higher intermediate pivot (higher I-pivot), as in example (1), or lower intermediate (lower I-pivot), as in (2), and VP-pivot, as in (3).[1]

There are restrictions on the quantifiers that can occur as determiners in these positions which have been studied in detail in Vangsnes (1999, 2002). Furthermore there are restrictions on the types of verbs that allow pivots in the different positions. Thráinsson (2007, p. 310 f.) gives examples of unaccusative and unergative intransitives, passives, middles, transitives and more. We will look first at unergative and transitive verbs. Unergative intransitive verbs of motion allow both I-pivots and VP-pivots whereas transitive verbs only allow I-pivots.

(4) *Það hafa* ***nokkrar rollur*** *hlaupið yfir veginn.*
EXPL have-PL some-NOM sheep-NOM run across road-the
'Some sheep have run across the road.'

(5) *Það hafa hlaupið* ***nokkrar rollur*** *yfir veginn.*
EXPL have-PL run some-NOM sheep-NOM across road-the
'Some sheep have run across the road.'

(6) *Það hefur* ***einhver*** *stolið hjólinu mínu.*
EXPL has somebody-NOM stolen bike mine
'Somebody has stolen my bike.'

1 Note that the higher I-pivot position, immediate after the tensed verb, can be argued to be a canonical subject position. This paper focuses on VP-pivots and we will not discuss whether one should distinguish between the higher I-pivot position and the canonical subject position.

(7) *Það hefur stolið **einhver** hjólinu mínu.
EXPL has stolen somebody-NOM bike mine
Intended: 'Somebody has stolen my bike.'

The same pattern seems to obtain with verbs that take an infinitival VP complement; I-pivots are possible, as in (8), but not VP-pivots. This is illustrated with a control verb in (9).

(8) Það hafa **margir** reynt að klífa fjallið.
EXPL have many-NOM tried to climb the-mountain
'Many people have tried to climb the mountain.'

(9) *Það hafa reynt **margir** að klífa fjallið.
EXPL have tried many-NOM to climb the-mountain
Intended: 'Many people have tried to climb the mountain.'

In the previous examples, the initial position is occupied by an expletive (*það*). As is well known, the expletive is restricted to clause-initial position in Icelandic, unlike in the mainland Scandinavian languages. When the tensed verb is in first position, as in yes/no questions, or when a non-subject is topicalized, no expletive shows up, as illustrated in the following examples; compare (10) with (1) and (11) with (6).

(10) Hafði (*það) **einhver** **köttur** alltaf verið í eldhúsinu?
had EXPL some-NOM cat-NOM always been in kitchen-the
'Had there always been some cat in the kitchen?'

(11) Auk þess hefur (*það) **einhver** stolið hjólinu mínu.
as-well-as this has EXPL somebody-NOM stolen bike mine
'In addition somebody has stolen my bike.'

In the examples given so far, the case of the pivot is nominative, but other cases are possible. The generalization is that the case of the pivot is the same as it would have been in a canonical subject position. The verb *reka* takes an accusative subject and the pivot is therefore accusative.

(12) **Nokkra** **hvali** hefur rekið á land í nótt.
several-ACC whales-ACC has driven to land in night
'Several whales have stranded overnight.'

(13) Það hefur rekið **nokkra** **hvali** á land í nótt.
EXPL has driven several-ACC whales-ACC to land in night
'Several whales have stranded overnight.'

We now turn to investigating the behavioral properties of the pivots, starting with reflexivization. Whereas objects can in some cases optionally control reflexives, subject control is obligatory in Icelandic. This is what we find in sentences such as (14) and (15) with I-pivots.

(14) *Það hafa* ***fjórir stúdentar*** *týnt hjólunum* ***sínum****/*þeirra.*
EXPL have four-NOM students-NOM lost bicycles-the their-REFL/*their-NON-REFL
'Four students have lost their bikes.'

(15) *Það hafa aldrei* ***fjórir stúdentar*** *týnt hjólunum* ***sínum****/*þeirra.*
EXPL have never four-NOM students-NOM lost bicycles-the their-REFL/*their-NON-REFL
'Four students have never lost their bikes.'

VP-pivots also control reflexives as shown in the following examples with the unaccusative verb *koma*.[2]

(16) *Það hafa* ***margir furðufuglar*** *komið hingað í dag með einkennilegar uppfinningar* ***sínar****/*þeirra.*
EXPL have many-NOM strange-fellows-NOM come here to day with peculiar inventions their-REFL/*their-NON-REFL
'Many strange fellows have come here today with their peculiar inventions.'

(17) *Það hafa komið* ***margir furðufuglar*** *hingað í dag með einkennilegar uppfinningar* ***sínar****/*þeirra.*
EXPL have come many-NOM strange-fellows-NOM here to day with peculiar inventions their-REFL/*their-NON-REFL
'Many strange fellows have come here today with their peculiar inventions.'

As the examples show, the reflexivization facts remain the same regardless of the position of the pivot.

The next test concerns subject ellipsis. An active clause with a VP-pivot may be coordinated with a subjectless clause, as shown in (18), provided that the tensed auxiliary is also omitted. This is not possible if the indefinite NP is an object of a transitive verb; then both an overt subject pronoun and a finite verb are required as shown in (19).

2 Examples (16) and (17) are adapted from Rögnvaldsson (1983).

(18) *Það hafa komið* ***margir furðufuglar*** *hingað í dag og*
EXPL have come many-NOM strange-fellows-NOM here to day and
farið í kröfugönguna.
gone to demonstration-the.
'Many strange fellows have come here today and gone to the demonstration.'

(19) *Við höfum hitt* ***marga furðufugla*** *og *(þeir hafa) farið*
we have met many-ACC strange-fellows-ACC and they have gone
í kröfugönguna.
to demonstration-the.
Intended: 'We have met many strange fellows and they have gone to the demonstration.'

The pattern shown in (18) has been referred to as 'pseudo-coordination' as it differs in many respects from ordinary coordination (see e.g. Wiklund 2007; Lødrup 2002; Kinn to appear). For our purpose, the label is not important; the difference in grammaticality between (18) and (19) shows that we need to make a distinction between postverbal pivots and objects.[3]

Zaenen et al. (1985) also show that, regardless of case, the understood subject argument of an embedded infinitival clause may be controlled by a subject, or object, in the matrix clause. The verb *vanta* 'to lack' takes both an accusative subject and an accusative object, see (20). The subject argument may be controlled as shown in (21) from Zaenen et al. (1985):454.

(20) *Mig vantar peninga.*
me-ACC lacks money-ACC
'I lack money'

(21) *Ég vonast til að vanta ekki peninga.*
I hope for to lack not money-ACC
'I hope not to lack money.'

However, the infinitival complement of a control verb is not a position in which we expect to find a presentational construction: the pivot would have to be coreferent with the subject or object of the matrix clause. In that case it would no longer be new information, so it does not fullfill the requirements for a presentational construction. Consequently this test is inapplicable to pivots.

The test for subject-to-object raising (also known as Exceptional Case Marking or Accusativus cum Infinitivo), however, reveals some interesting facts. In addition to the expected version in (22), where a subject in canonical position 'exceptionally' receives accusative case (Thráinsson 2007, p. 149), the word order in (23) is also possible.

3 Lødrup (2002, p. 123) actually argues that subject ellipsis is ungrammatical in presentational sentences but his example does not involve pseudo-coordination and has an overt finite verb in the second clause.

(22) *Jón telur* ***hestana*** *hafa verið í kirkjugarðinum.*
John believes horses-the-ACC to-have been in churchyard-the
'John believes the horses to have been in the churchyard.'

(23) *Jón telur (*það) hafa verið* ***hesta*** *í kirkjugarðinum.*
John believes there to-have been horses-ACC in churchyard-the
'John believes there to have been horses in the churchyard.'

As expected, there is no expletive in the embedded clause, but the post-verbal position of the indefinite *hesta* suggests that this is a presentational structure, as indicated in the paraphrase. It has acquired the accusative case we would expect in an AcI construction, not the nominative, which we would expect when the case is not lexically assigned.[4] So the pivot seems to have been raised. The situation can be seen as similar to that of backwards raising or control (as discussed in e.g. Polinsky and Potsdam 2012).

There are also passive versions such as (24), or even, although less good, with subject-to-subject raising, as in (25).

(24) *Það voru taldir vera* ***hestar*** *í kirkjugarðinum*
EXPL were believed-MASC.PL to-be horses-NOM-MASC.PL in churchyard-the
'There were believed to be horses in the churchyard.'

(25) *?Það voru* ***hestar*** *taldir vera í kirkjugarðinum*
EXPL were horses-NOM-MASC.PL believed-MASC.PL to-be in churchyard-the
'There were horses believed to be in the churchyard.'

Notice that in the first of these passives, the matrix verb agrees with the postverbal nominative in the embedded infinitive. We assume a raising analyis for these constructions, but their analysis seems to be very much in flux (see Thráinsson 2007, pp. 452–458 for some discussion).

To summarize, not all tests for subject properties that were used for canonical subjects in Zaenen et al. (1985) are applicable to pivots in Icelandic. But the ones that can be used (reflexivization, subject ellipsis and, arguably, raising) show that pivots behave like subjects.

4 There is evidence from adjuncts that PRO in Icelandic has the case an overt subject would have in a finite clause (Sigurðsson 1991).

2.2 Swedish

With respect to presentational constructions, Swedish differs from Icelandic in two ways. First, I-pivots are not possible, only VP-pivots. Compare the Swedish version of the Icelandic examples in (1)–(3) shown in (26)–(28).

(26) **Det hade **en katt** alltid varit i köket.*
EXPL had a cat always been in kitchen-the
'There had always been a cat in the kitchen.'

(27) **Det hade alltid **en katt** varit i köket.*
EXPL had always a cat been in kitchen-the
'There had always been a cat in the kitchen.'

(28) *Det hade alltid varit **en katt** i köket.*
EXPL had always been a cat in kitchen-the
'There had always been a cat in the kitchen.'

In periphrastic passive clauses, the pivot typically appears after the auxiliary but in front of a participle which agrees with the pivot.[5]

(29) *Det hade blivit **så många studenter** antagna.*
EXPL had become so many students-PL admitted-PL
'There had been so many students admitted.'

Second, the expletive subject is not limited to initial position, but may also occur after the finite verb, e.g. in questions, see (30).

(30) *Hade det alltid varit **några katter** i köket?*
had EXPL always been some cats in the-kitchen
'Had there always been some cats in the kitchen?'

Presentational sentences with transitive action verbs (31) and control verbs (32) are impossible, as in Icelandic (7) and (9).[6]

(31) **Det har stulit **någon student** cykeln.*
EXPL has stolen some student bike-the
Intended: 'Some student has stolen the bike.'

5 In Danish and Norwegian, the pivot normally follows the participle in such constructions, see Engdahl and Laanemets (2015) and Engdahl (2017). See also Holmberg (2002) for a comparison with Icelandic.

6 In earlier stages, Swedish appears to have been more like Icelandic, allowing I-pivots with transitive verbs (see Håkansson 2017).

(32) **Det har försökt* ***många*** *att bestiga berget.*
EXPL have tried many-NOM to climb mountain-the
Intended: 'Many people have tried to climb the mountain.'

As for case marking, since only pronouns show case in Swedish and personal pronouns are normally not possible in presentational constructions, we wouldn't expect case to show up on the pivot. There is however one construction that allows for a personal pronoun and this can only have nominative case, see (33) from Teleman et al. (1999, Vol. 3, p. 387). The definite pronoun *de* together with a relative clause gets a kind reading.

(33) *Det lär finnas* ***de*** *som fortfarande stöder regeringen.*
EXPL MOD exist they-NOM that still support government-the
'There are supposed to be people who still support the government.'

In Swedish, as in Icelandic, clause-internal pronominalization under identity with a subject requires a reflexive, regardless of whether the subject is in canonical position or a VP-pivot, see (34), adapted from Börjars and Vincent (2005).

(34) *Det hade kommit* ***en man*** *med* ***sin****/*hans fru.*
EXPL had come a man with his-REFL/his-NON-REFL wife
'There had come a man with his (own) wife.'

With respect to subject ellipsis, active clauses with VP-pivots may be pseudo-coordinated, as observed in Börjars and Vincent (2005) and Engdahl (2006). As in Icelandic, the coordinated verbs must agree in tense and auxiliaries are not repeated. This type of coordination is not possible with objects, see (36).

(35) *Det har kommit* ***en student*** *och frågat efter dig.*
EXPL has come-SUP a student and asked-SUP after you
'A student has come and asked for you.'

(36) *Vi har träffat* ***några studenter*** *och *(de har) frågat efter dig.*
we have met some students and they have asked-SUP after you
'We have met some students and they have asked about you.'

As for the raising-to-object test, the only argument that may raise in Swedish is the overt expletive which is generated in canonical subject position. A Swedish version of the Icelandic example (23) is given in (37).

(37) *Johan anser det ha varit* ***för många hästar*** *på kyrkogården.*
John considers EXPL have been too many horses on churchyard-the
'Johan considers there to have been too many horses in the churchyard.'

In addition we find examples like (38) where the expletive is a canonical subject of a passive matrix verb. However, (38) is probably best seen as an impersonal passive given that inserting an overt agent phrase such as *av Johan* 'by Johan' is infelicitous.

(38) *Det anses ha varit* ***för många hästar*** *på kyrkogården.*
EXPL consider-PASS have been too many horses on churchyard-the
'It is believed that there have been too many horses in the churchyard.'

Unlike Icelandic, the case of the pivot remains nominative in Swedish. The following example is somewhat stilted, but the pronoun has to be nominative.

(39) *Johan anser det omöjligen kunna finnas* ***de*** *som*
John considers EXPL impossibly can-INF exist they-NOM that
tror att jorden är platt.
believe that earth-the is flat
'Johan considers it impossible that there exist people who believe that the earth is flat.'

We conclude that the reflexivization and subject ellipsis tests show that pivots in Swedish also have syntactic subject properties. But in the AcI construction we see that the expletive also has a syntactic subject property.

2.3 What identifies subjects in Scandinavian languages?

It has emerged from the previous discussion that in Icelandic VP-pivots are grammatical subjects under the criteria proposed in Zaenen et al. (1985), whereas the expletive has no subject properties. This leads to the somewhat paradoxical conclusion that in Icelandic, neither case marking nor position uniquely identify subjects. Following Zaenen et al. (1985), Sigurðsson (2004) and others, it seems to have been assumed that position was the relevant coding property since case marking didn't work, but the facts above suggest that this is not generally true. Nor is it an either/or condition, since we can find 'quirky' VP-pivots which also control reflexives, as shown in (40).

(40) *Það hefur að sögn rekið* ***nokkra hvali*** *á land í nótt*
EXPL has to report driven several-ACC whales-ACC to land in night
með kálfum sínum.
with calves their-REFL.
'Reportedly several whales have stranded overnight with their calves.'

In Swedish, the situation is more complicated; reflexivization and pseudo-coordination give the same result as in Icelandic: the pivot behaves as a subject. But the expletive undeniably behaves as a subject in terms of position and raising.[7]

7 This is reflected in the terminology used in the reference grammars where both the expletive and the pivot are referred to as subjects. The expletive is commonly referred to as *formellt subjekt* 'formal subject'. Teleman et al. (1999) refers to the pivot as *egentligt subjekt* 'real subject' and Faarlund et al. (1997) use the term *potensielt subjekt* 'potential subject'.

What is then the theoretical status of the canonical subject positions? They are clearly the statistically most prevalent positions in which subjects are found in Icelandic and Swedish, but that is hardly a syntactic distinction. They can also be claimed to be unmarked positions in the sense that all types of subjects can occur in these positions, whereas the other positions are more restricted. But bare non-specific indefinites are, in fact, not very good in the canonical positions. Thráinsson (2007, p.323) gives a question mark to (41).

(41) *?Mús* *hefur verið í* *baðkerinu.*
mouse-NOM has been in bathtub-the
'A mouse has been in the bathtub.'

An indefinite article is required in the corresponding Swedish example in (42), which is grammatical, but somewhat marked compared to a presentational construction.

(42) *En mus* *har varit i* *badkaret.*
a mouse has been in bathtub-the
'A mouse has been in the bathtub.'

Whether these facts are interesting from a syntactic point of view depends on the nature of these constraints: if, as has often been claimed, they are pragmatic in nature (e.g. based on discourse structure), it is not immediately clear that they should be accounted for in syntactic terms.

In the next section we discuss some of the constraints that have been proposed on VP-pivots. While we will not be able to elucidate the nature of these constraints completely, we hope to at least present enough data to provide a good basis for a more substantial study.

3 Constraints on VP-pivots

The findings in the previous section go against a widely held belief that the indefinite NP in presentational sentences in Scandinavian languages is an object (see e.g. Lødrup 1999). But it is not the case that any indefinite subject can occur in the non-canonical positions. As shown in Vangsnes (1999, 2002) there are constraints on which quantifiers are possible, summarized in Thráinsson (2007). Another source of constraints is the thematic relation between the verb and its subject argument. These were first discussed in Platzack (1983), who assumed that what we are here calling I-pivots and VP-pivots are generated in different positions, I-pivots outside the VP and VP-pivots inside the VP. In addition he proposed a correlation between syntactic positions and the types of theta roles that can be generated there.[8] Maling (1988) elaborated on Platzack's analysis and argued that grammatical rules need to refer both to thematic roles and to the

8 In later work, Platzack (2010) has made this connection explicit, referring to the *Uniformity of Theta Assignment Hypothesis* (UTAH) of Baker (2006).

mapping between the thematic hierarchy and the syntactic hierarchy. Lødrup (1999) assumed that the VP-pivot is an object, albeit an atypical one since it may have agentive properties, whereas Faarlund et al. (1997) point out certain differences between VP-pivots of active sentences and objects. In this section we take a closer look at the interaction between position and thematic properties.

3.1 Icelandic

In Icelandic, subjects of all lexical semantic verb types seem to be possible as I-pivots but not as VP-pivots. We have already seen that the agent argument of a typical active transitive verb cannot occur inside the VP, cf. (7). However, this does not seem to be linked to the transitivity of the verb, as proposed in the analysis by Platzack (1983), since our informants prefer I-pivots also with intransitive verbs with agentive subjects like *hringa* 'phone', as shown in (43) and (44).

(43) *Það hafði* ***margt fólk*** *hringt í mig í gær.*
EXPL had many-NOM people-NOM phoned to me on yesterday
'Many people had phoned me yesterday.'

(44) *?Það hafði hringt* ***margt fólk*** *í mig í gær.*
EXPL had phoned many-NOM people-NOM to me on yesterday
'Many people had phoned me yesterday.'

Similarly, experiencer arguments are acceptable as I-pivots, but not as VP-pivots.

(45) *Það hefur* ***mörgum börnum*** *verið kalt.*
EXPL have many-DAT children-DAT been cold.
'Many children have been cold.'

(46) **Það hefur verið* ***mörgum börnum*** *kalt.*
EXPL have been many-DAT children-DAT cold.
'Many children have been cold.'

The goal or recipient argument of *hjálpa* 'help' is fine as an I-pivot, but not as a VP-pivot.

(47) *Það var* ***gömlum manni*** *hjálpað yfir götuna.*
EXPL was old-DAT man-DAT helped across street-the
'An old man was helped across the street.'

(48) *?*Það var hjálpað* ***gömlum manni*** *yfir götuna.*
EXPL was helped old-DAT man-DAT across street-the
'An old man was helped across the street.'

Maling (1988) shows that it is not the case marking but the thematic role that is relevant in these examples. What seems to be at issue is how thematic roles are mapped onto syntactic positions.[9] Maling demonstrates that whereas it is impossible to realize indefinite experiencer subjects as VP-pivots, as shown in (49), it is possible to find theme subjects with the few verbs that have a theme subject and an experiencer object, as in (50).

(49) **Það hafa óttast* ***margir lögreglumenn*** *fjölgun slysa.*
EXPL have feared many-NOM police-officers-NOM increase-ACC accidents-GEN
Intended: 'Many police officers feared an increase in accidents.'

(50) *Það hefur hraett* ***einhver mynd*** *börnin.*
EXPL has frightened some-NOM picture-NOM children-the-ACC
'Some picture has frightened the children.'

It is clear then that the constraint is not against having two NPs in the VP — given the existence of ditransitive verbs in Icelandic such a constraint would be rather astonishing — but needs to be stated in semantic/thematic terms. We can summarize the findings for Icelandic as follows: I-pivots can occur with all kinds of thematic roles but VP-pivots are only possible with themes. A more precise statement of the constraints, however, needs further research.

Faarlund et al. (1997, p. 846 f.) claim that in Norwegian VP-pivots in active clauses behave differently from VP-pivots in passive clauses. Using reflexivization and coordination tests, they show that only the pivots in active clauses have the typical subject properties identified in Section 2. However, applying their tests to Icelandic gives somewhat different results.[10] The VP-pivot of a passive verb still controls reflexives, as shown in (51).

(51) *Það var fleygt* ***nokkrum stúdentum*** *út af skrifstofum sínum/??þeirra.*
EXPL was kicked some students out of office their-REFL/NON-REFL
'Some students were kicked out of their offices.'

9 Examples (45)–(50) are from Maling (1988).

10 Faarlund et al. (1997) use a third test involving control of adjuncts. We found that while this distinguishes between canonical subjects and pivots in passive clauses, it did not reliably distinguish between pivots in active and passive clauses.

As for coordination, a passive clause with a VP-pivot may be conjoined with a passive VP, without subject and auxiliary, as in (52). Subject ellipsis in (53) with an active VP in the second conjunct is ungrammatical.[11]

(52) *Það hafa verið seldir* ***margir bílar*** *og*
EXPL have been sold-MASC.PL many-NOM cars-NOM-MASC.PL and
fluttir út til Póllands.
exported-MASC.PL out to Poland
'There have been many cars sold and exported to Poland.'

(53) *Það var fleygt* ***nokkrum stúdentum*** *út af skemmtistaðnum*
EXPL was kicked several-DAT students-DAT out of nightclub-the
*og *(þeir) urðu æstir.*
and they were upset
Intended: 'Several students were kicked out of the nightclub and they were upset.'

In Icelandic, VP-pivots in passive clauses thus show mixed properties. They control reflexivization, like canonical subjects, but are less acceptable in coordination than VP-pivots in active clauses. In addition there is an interaction between thematic roles and the passive, as shown in Maling (1988).

3.2 Swedish

We will distinguish between intransitive and transitive constructions. We first note that, with intransitive predicates, Swedish, unlike Icelandic, allows VP-pivots with verbs that normally are interpreted as having agentive subjects as described by e.g. Anward (1981) and Teleman et al. (1999, Vol. 3, p. 400 f.).

(54) *Det brukade arbeta* ***många människor*** *här.*
EXPL used-to work many people here
'Many people used to work here.'

(55) *Det har sjungit* ***några islänningar*** *i vår kör.*
EXPL have sung some Icelanders in our choir
'Some Icelanders have sung in our choir.'

Anward (1981, p. 10) points out that the activity meaning tends to fade away and that the location of the activity is foregrounded when these verbs are used in presentational sentences. He cites as evidence the fact that adding an intentional subject-oriented adverb is infelicitous, see (56). According to Teleman et al. (1999, Vol. 3, p. 400), the

11 Both these examples involve ordinary coordination; see also Eythórsson (2008, p. 179 f.), who discusses similar examples.

verbs tend to denote activities which are typical in some location or context, such as singing in a choir. They note that it would be strange to emphasize the manner, see (57).

(56) *Det har (*motvilligt) arbetat* ***många människor*** *(*motvilligt) här.*
EXPL has reluctantly worked many people reluctantly here
Intended: 'Many people have reluctantly worked here.'

(57) *?Det har sjungit* ***några islänningar*** *entusiastiskt i vår kör.*
EXPL have sung many Icelanders enthusiastically in our choir
Intended: 'Many Icelanders have sung enthusiastically in our choir.'

However, verbs like *ringa* 'phone', which don't seem to require a location in the presentational construction, can also be used, as shown in (58).

(58) *Det har ringt* ***nån*** *till dig.*
EXPL has phoned someone to you
'Someone has phoned you.'

But here too the focus seems to be on the event, that there was a phone call, not on the agentivity of the caller.[12] Recall that our Icelandic informants prefer the I-pivot version of this example, see (43) and (44), but this option is of course not available in Swedish.

Example (59), adapted from Maling (1988), shows that with intransitive verbs, an experiencer argument cannot be realized as a VP-pivot, which we have seen is impossible also in Icelandic, see (45)–(48).

(59) **Det hade frusit* ***några barn*** *i natt.*
EXPL had frozen some children in night
Intended: 'Some children had felt cold last night.'

We note in passing that verbs taking experiencer subjects are fine when pseudo-coordinated with a presentational clause, see (60).

(60) *Det hade suttit* ***några barn*** *utanför och frusit.*
EXPL had sat some children outside and frozen
'Some children had sat outside and felt cold.'

This suggests that whatever the constraint against indefinite experiencers is, it only applies to VP-pivots. Once such an indefinite NP has been introduced in the first conjunct, it seems to provide an antecedent for subject ellipsis in the second conjunct.[13]

12 Lødrup (2002, p. 122) notes that communication verbs like *ringe* 'call' are fine in Norwegian presentational constructions.

13 We now have an explanation for the observation made in Engdahl (2006, p. 41), viz. that it is possible to add an adverb like *motvilligt* 'reluctantly' in a follow-up clause to a presentational sentence with *arbeta* 'work'. This is because the presentational sentence introduces a referent which can be referred to in a later clause, essentially the same explanation as for why experiencer verbs are possible in second conjuncts, as in (60).

Passive verbs allow a goal subject to be realized as a VP-pivot in Swedish, see (61), unlike Icelandic where only the I-pivot is acceptable, as shown in (47) and (48).

(61) *Det har hjälpts* ***tusentals flyktingar*** *i det här lägret.*
EXPL have help-PASS thousands refugees in this camp-the
'Thousands of refugees have been helped in this camp.'

But experiencer subjects of passive verbs are unacceptable as VP-pivots.

(62) *?*Det har skrämts* ***många barn*** *med berättelser om tomten.*
EXPL has frightened-PASS many children with stories about Santa-Claus-the
Intended: 'Many children have been frightened with stories about Santa Claus.'

As already observed in Maling (1988, p. 180), there is a difference between Icelandic and Swedish regarding the mapping between thematic roles and syntactic positions: Icelandic has a choice between I-pivots and VP-pivots. Agents, goals and experiencers, which are unacceptable as VP-pivots, are fine as I-pivots in that language. Swedish, having only one pivot position, seems to relax the thematic constraint so that agents can fill this position in intransitive actives and goals in passives, whereas experiencers are unacceptable.

In transitive constructions, as we already mentioned, Swedish does not allow VP-pivots with agentive verbs like *steal*, as shown in (31), repeated here as (63).

(63) **Det har stulit* ***någon student*** *cykeln.*
EXPL has stolen some student bike-the
Intended: 'Some student has stolen the bike.'

We do, however, find presentational sentences with two NPs inside the VP, as already pointed out in Platzack (1983). The following examples are adapted from his article.[14] In these examples, the pivot is clearly non-agentive, arguably a theme.

(64) *Det hade hänt honom* ***något konstigt*** *igår.*
EXPL had happened him something strange yesterday
'Something strange had happened to him yesterday.'

(65) *Det kunde vänta mig* ***en verklig överraskning*** *när jag kom hem.*
EXPL could await me a real surprise when I came home
'A real surprise could be waiting for me when I came home.'

14 Platzack's examples have single finite verbs and could be analyzed as involving some form of object shift, as pointed out by an anonymous reviewer.

Passive versions of ditranstive verbs provide another context where there is more than one NP inside the VP. Example (66) is also adapted from Platzack (1983). Note that the VP-pivot can only realize a theme argument, not a goal argument.

(66) *Det hade tilldelats studenten* ***en belöning****.*
EXPL had given-PASS student-the an award
'The student had been given an award.'

(67) **Det hade tilldelats* ***en student*** *belöningen.*
EXPL had given-PASS a student award-the
Intended: 'The award had been given to a student.'

However, we don't find any good Swedish counterparts to the Icelandic theme–experiencer example in (50) despite the possibility of examples like (64)–(65).

(68) **Det hade skrämt barnen* ***nån bild****.*
EXPL had frightened children-the some picture
Intended: 'Some picture had frightened the children.'

It may be that (68) is impossible because the Swedish verb *skrämma* 'frighten' is more strongly agentive, or causative, than e.g. *hända* 'happen', since *skrämma* is also used with animate subjects, unlike *hända*. So in Swedish too, only theme VP-pivots are possible when there is another NP argument in the VP.

We now turn to the possible syntactic differences between VP-pivots in active and passive clauses. As for reflexivization, the overall pattern is the same as in Icelandic with VP-pivots preferably controlling reflexive pronouns in passive clauses (69), but there seems to be more variation in Swedish than in Icelandic (cf. Teleman et al. 1999, Vol. 3, p. 394). The opposite preference shows up when the antecedent is an ordinary object, as in (70).

(69) *Det hade körts ut* ***några studenter*** *från sina/?deras*
EXPL had kick-PASS out some students from their-REFL/NON-REFL
kontor.
offices
'There had been some students kicked out of their offices.'

(70) *Man hade kört ut* ***några studenter*** *från deras/?sina*
Someone had kicked out some students from their-NON-REFL/REFL
kontor.
offices
'Someone had kicked out some students from their offices.'

The coordination test also gives the same result for Swedish as for Icelandic. Coordination of two passive VPs is possible, see (71), but when the second conjunct is active, an overt subject pronoun is required as shown in (72).

(71) *Det har sålts* ***många bilar*** *och exporterats till Polen.*
EXPL has sold-PASS many cars and exported-PASS to Poland
'There had been many cars sold and exported to Poland.'

(72) *Det hade körts ut* ***några studenter*** *och *(de) var upprörda.*
EXPL had kicked-PASS out some students and they were upset
'There had been some students kicked out and they were upset.'

Swedish then is similar to Icelandic in that (typical) experiencers are not realized as VP-pivots. As for agents, we find two differences. In Icelandic, agents of transitive verbs can be realized as I-pivots, but this option is not available in Swedish. Agent-like arguments of intransitives are acceptable as VP-pivots in Swedish, but there is a constraint against subject-oriented intentional adverbs and manner adverbs. This constraint suggests that the agentivity of the subject argument is somehow reduced in the presentational construction. It is, however, difficult to pin down what exactly that means. It is unlikely that these agents cannot be seen as having volition; it seems more plausible that the construction does not single out the pivot itself but instead introduces an event, or a situation, as a whole.[15] Our investigation also confirms that in Swedish, as in Icelandic, the constraint is not on the number of positions in the VP but rather on which thematic roles can be realized there.

3.3 Position versus thematic roles

In previous sections we have shown that both position and thematic roles matter when it comes to accounting for what subject properties the pivots in presentational sentences have. In Icelandic, we need to distinguish I-pivot positions from VP-pivot positions since there are more restrictions on the latter. For instance, subjects of transitive verbs cannot occur there, see (7), nor can goals or experiencers, regardless of whether the verb is intransitive, see (45) and (46), or transitive, see (49). In Swedish the intermediate non-canonical subject positions are not available, see (26)–(28). Agentive intransitives are possible but agentive transitive verbs are excluded, see (31) and (32). We find partly similar thematic restrictions on subjects inside the VP as in Icelandic; experiencer pivots are excluded but goals are possible.

Whereas VP-pivots of active verbs behave much like subjects in canonical positions — they control reflexives and allow subject ellipsis in a pseudo-coordinated VP — VP-pivots in passive clauses control reflexives but don't allow subject ellipsis, see (53) and

15 See Sveen (1996) for extensive discussion of similar facts in Norwegian.

(72). In this respect they behave more like ordinary objects. This constraint on passives might come as a surprise. But a bit of reflection makes it less surprising: the pivot in the passive case is not the 'logical' subject. Passivization is an argument promotion operation, whereas the presentational construction demotes that same argument. The passivization strategy in the presentational sentences ends up demoting an argument which has already been promoted. It seems that this Duke of York gambit meets with ambivalence in the Scandinavian languages. While this might make intuitive sense, further study is needed of the conditions on both the passive and the presentational construction and of the mechanics that would make such a constraint on the argument mapping possible.

4 Conclusion

In this paper we have discussed whether VP-pivots in Icelandic and Swedish have syntactic subject properties. The only explicit discussion of a similar topic that comes to mind is that in Bonami et al. (1999) who discuss Stylistic Inversion for French and contrast a set of subject and object properties for the post-verbal NPs in that language. We have shown that the status of the indefinite NP in presentational constructions in Scandinavian languages is less clear than has been claimed in the literature about Norwegian. In both Swedish and Icelandic, these NPs have syntactic subject properties, even when they occur in VP-complement positions. In Icelandic, we find a rather neat partition of the subject properties that Keenan (1976) called coding properties and behavioral properties: only canonical subjects have the positional coding properties, whereas pivots share the behavioral properties with them. This brings to mind observations made by several authors (see e.g. Lambrecht 1994, pp. 131–145), that subjects tend to be unmarked topics. Under this view, the positional coding properties are actually properties of topics, not of subjects per se.

Present-day Swedish differs from Icelandic in having an expletive that clearly has the same coding properties as canonical subjects. The expletive also behaves like a subject in subject to object raising. So, in this language, there is no neat line-up of the properties following Keenan's (1976) classification together with the hypothesis that the positional properties are topic properties. However, in earlier stages of Swedish, the position of the expletive was more similar to the situation in present-day Icelandic (see Håkansson 2017), which suggests that one should look at the diachronic development as well. One further similarity between Icelandic and Swedish is that the VP-pivot shows nominative case, see (33).[16] This distinguishes Icelandic and Swedish from Danish and Norwegian where the pivot has been claimed to be accusative (see Mikkelsen 2002; Lødrup 1999). It remains to be seen whether this morphological difference correlates with differences in the syntactic subject properties that are the topic of this paper.

16 Unless the Icelandic verb has a lexically selected case as in (13).

Acknowledgments

For the Icelandic data in this paper we have consulted four native speakers. We thank Einar Freyr Sigurðsson, Halldór Ármann Sigurðsson, Jóhannes Gísli Jónsson and Sigríður Sigurjónsdóttir for their help. Some Swedish judgments were checked with five other speakers and the remainder reflect the intuitions of the Swedish co-author. We also acknowledge the comments from three anonymous reviewers.

References

Anward, Jan (1981). *Functions of passive and impersonal constructions. A case study from Swedish*. Ph.D. dissertation, Uppsala University.

Askedal, John Ole (1986). "Ergativity in Modern Norwegian". In: *Nordic Journal of Linguistics* 9, pp. 25–45.

Baker, Mark (2006). "Handbook in generative syntax: Elements of grammar". In: *Thematic roles and syntactic structures*. Ed. by Liliane Haegeman. Dordrecht: Kluwer, pp. 73–137.

Beaver, David, Itamar Francez, and Dmitry Levinson (2005). "Bad subject: (Non-)canonicality of NP distribution in Existentials". In: *SALT XV*, pp. 19–43.

Bonami, Olivier, Danièle Godard, and Jean-Marie Marandin (1999). "Constituency and word order in French subject inversion". In: *Constraints and resources in natural language syntax and semantics*. Ed. by G. Bouma, E. Hinrichs, G.-J. Kruijff, and R. Oehrle, pp. 21–40.

Börjars, Kersti and Nigel Vincent (2005). "Position versus function in Scandinavian presentational constructions". In: *Proceedings of the LFG05 Conference*. Ed. by Miriam Butt and Tracy King. CSLI, pp. 54–72.

Engdahl, Elisabet (2006). "Semantic and syntactic patterns in Swedish passives". In: *Demoting the agent: Passive, middle and other voice phenomena*. Ed. by Benjamin Lyngfelt and Torgrim Solstad. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 21–45.

– (2017). "Expletive passives in Scandinavian – with and without objects". In: *Order and structure in syntax II: Subjecthood and argument structure*. Ed. by Laura Bailey and Michelle Sheehan. Language Science Press, pp. 291–308.

Engdahl, Elisabet and Anu Laanemets (2015). "Opersonlig passiv i danska, norska och svenska – en korpusstudie". In: *Norsk Lingvistisk Tidsskrift* 33, pp. 129–483.

Eythórsson, Thórhallur (2008). "The New Passive in Icelandic really is a passive". In: *Grammatical Change and Linguistic Theory: The Rosendal Papers*. Ed. by Thórhallur Eythórsson. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 173–219.

Faarlund, Jan Terje, Svein Lie, and Kjell Ivar Vannebo (1997). *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Håkansson, David (2017). "Transitive expletive constructions in Swedish". In: *Nordic Journal of Linguistics* 40(3).

Holmberg, Anders (2002). "Expletives and Agreement in Scandinavian Passives". In: *Journal of Comparative Germanic Linguistics* 4, pp. 5–128.

Keenan, Edward (1976). "Towards a universal definition of subject". In: *Subject and Topic*. Ed. by Charles Li. Academic Press, pp. 303–333.

Kinn, Torodd (to appear). "Asymmetric verb phrase coordination in Norwegian. Degrees of grammaticalization and constructional variants". In: *Grammaticalization meets Construction Grammar*. Ed. by Evie Coussé, Peter Andersson, and Joel Olofsson. Amsterdam: John Benjamins.

Lambrecht, Knut (1994). *Information structure and sentence form*. Cambridge University Press.

Lødrup, Helge (1999). "Linking and Optimaility in the Norwegian Presentational Focus Construction". In: *Nordic Journal of Linguistics* 22, pp. 205–230.

– (2002). "The syntactic structures of Norwegian pseudocoordinations". In: *Studia Linguistica* 56, pp. 121–143.

Maling, Joan (1988). "Variations on a theme: Existential sentences in Swedish and Icelandic". In: *McGill Working Papers in Linguistics*, pp. 168–191.

Mikkelsen, Line (2002). "Reanalyzing the Definiteness Effect: Evidence from Danish". In: *Working Papers in Scandinavian Syntax* 69, pp. 1–75.

Platzack, Christer (1983). "Existential sentences in English, Swedish, German and Icelandic". In: *Papers from the seventh Scandinavian Conference of Linguistics*. Ed. by Fred Karlsson, pp. 80–100.

– (2010). *Den fantastiska grammatiken. En minimalistisk beskrivning av svenskan*. Stockholm: Norstedts.

Polinsky, Maria and Eric Potsdam (2012). "Backward Raising". In: *Syntax 15*, pp. 75–108.

Rögnvaldsson, Eiríkur (1983). "Sagnliðurinn í íslensku". In: *Íslenskt mál* 5, pp. 7–28.

Sigurðsson, Halldór Ármann (1991). "Icelandic case-marked PRO and the licensing of lexical arguments". In: *Natural Language and Linguistic Theory* 9, pp. 327–363.

– (2004). "Icelandic non-nominative subjects: Facts and implications". In: *Non-nominative Subjects Vol.2*. Ed. by P. Bhaskararao and K.V. Subbarao, pp. 137–159.

Sveen, Andreas (1996). *Norwegian Impersonal Actives and the Unaccusative Hypothesis*. Dr.art. thesis, University of Oslo.

Teleman, Ulf, Staffan Hellberg, and Erik Andersson (1999). *Svenska Akademiens grammatik*. Stockholm: Norstedts.

Thráinsson, Höskuldur (2007). *The Syntax of Icelandic*. Cambridge: Cambridge University Press.

Vangsnes, Øystein (1999). *The identification of functional architecture*. Doctoral Dissertation, University of Bergen.

– (2002). "Icelandic expletive constructions and the distribution of subject types". In: *Subjects, Expletives, and the EPP*. Ed. by Peter Svenonius. Oxford: Oxford University Press, pp. 43–70.

Wiklund, Anna-Lena (2007). *The Syntax of Tenselessness*. Berlin: Mouton de Gruyter.

Zaenen, Annie, Joan Maling, and Höskuldur Thráinsson (1985). "Case and grammatical functions". In: *Natural Language and Linguistic Theory* 3.4, pp. 441–483.