Learner Corpus Research: LCR2013 Conference Proceedings

BeLLS 2015 Vol. 6, ISBN 978-82-998587-7-9, s. 1–4 BeLLS.uib.no

Preface

Ann-Kristin Helland Gujord[1], Susan Nacey[2] and Silje Ragnhildstveit[1]
[1]University of Bergen, [2]Hedmark University College

In September 2011, researchers from around the world gathered in Louvain-la-Neuve, Belgium, for the first Learner Corpus Research (LCR) conference. This conference was the initiative of scholars from the Centre for English Corpus Linguistics (CECL), which had been launched at the University of Louvain twenty years earlier and had pioneered the development and study of learner corpora under the direction of Sylviane Granger. In essence, the foundation of the CECL in the late 1980s marked the genesis of learner corpus research as a new field, merging elements from corpus research, foreign/second language acquisition and language teaching. In turn, LCR2011 *20 years of learner corpus research: Looking back, moving ahead,* marked a transitional point in the field – a look towards the past to review (and celebrate) the progress made, but a look towards the future and the challenges that lay ahead.
Two years later, LCR2013 was held in Bergen/Os (Norway) at the express invitation of Granger, making a biannual tradition out of the LCR conference. In many ways, the move from Louvain to Bergen represented a symbolic passing of the baton. While the University of Louvain had Granger, the University of Bergen had Kari Tenfjord, who was a driving force behind the compilation of a learner corpus of L2 Norwegian, the *Norsk Andrespråkskorpus* (ASK); Tenfjord went on to lead the *ASKeladden project – A corpus-based approach to L1 transfer in Norwegian learner language*. Such long-term dedication to learner corpus research and language transfer research made Bergen a natural home for LCR2013.

The conference was held with the joint support of the Department of Linguistic, Literary and Aesthetic Studies at the University of Bergen, and the Norwegian Research Council. It was a resounding success, with roughly 120 international delegates; 49 papers and 21 posters and demonstrations were presented over the three-day period. Research topics ranged from presentations of new corpora and compilation issues to methodological concerns and suggestions to analyses of different aspects of learner language. The papers included in this edition of BeLLS, all of which were presented at LCR2013, deal variously with these three areas. Together, they offer a glimpse into the directions learner corpus research has taken since the field's conception, as well as indications about new avenues for continued research. The eight articles in this volume represent the diversity of language combinations and corpora which has begun to characterize learner corpus research, as well as demonstrate the high quality of the conference.

Four of the eight articles demonstrate the role of learner corpus research in the study second language acquisition (SLA), exploring the potential role of L1 influence in the acquisition and use of a second language: *Nicolai Struc and Nicholas Wood* report from a study of lexical transfer in written texts by Japanese learners of English extracted from a

Bergen Language and Linguistic Studies (BeLLS), May 30th 2015. © Gujord, Nacey and Ragnhildstveit
DOI: 10.15845/bells.v6i0.804

longitudinal learner corpus and a native-speaker corpus in "**Crosslinguistic lexical transfer of English-based loanwords in English L2 writing by Japanese university students**". They observe that Japanese learners rely on English-based loanword cognate items to a larger extent than native speakers when writing argumentative and narrative texts in English. Struc and Wood link this finding to the considerable number of loanwords in the Japanese lexicon - so-called 'gairaigo' - and draw into question previous studies suggesting that Japanese knowledge and awareness of gairaigo can facilitate lexical acquisition in the L2. Methodological advances have recently been given much attention in transfer research, and this development is reflected in "**Tracing crosslinguistic influences in structural sequences: What does key structure analysis have to offer**?" by *Ilmari Ivaska.* Whereas Struc and Woods' transfer study may be accounted for within the 'comparison-based' approach, Ivaska demonstrates how transfer can be identified based on a 'detection-based' approach. In the detection-based approach, transfer effects are analyzed in data-driven manner without pre-existing hypotheses, and without comparisons of languages and interlanguage performances. Linguistic features of L2 that distinguish writers of different L1 backgrounds are automatically detected, commonly by means of statistical tools comparing frequencies of linguistic features in a text sample. Ivaska explores how a specific methodological procedure, key structure analysis, may serve as a methodological tool for automatically identifying the L1-specific behaviour of L2 Finnish texts retrieved from a learner corpus (The Corpus of Advanced Learner Finnish) written by speakers of five different L1-backgrounds. In a step-way approach Ivaska compares frequencies of lexical bundles of different length, and finds that 1-grams with conjunctions or present tense singular verbs, and 2-grams with a singular nominative and plural partitive distinguish well different l1 backgrounds in the current data set. L1 influence is also one of the topics in *Susan Nacey and Anne-Line Graedler's* study of inappropriate preposition uses in oral recordings by Norwegian advanced learners of English (LINDSEI–NO), "**Preposition use in oral and written learner language**". According to Nacey and Graedler, nearly half of the inappropriate uses can be accounted for in terms of negative transfer, particularly; the Norwegian learners have trouble distributing the preposition *on* in a target like manner. Furthermore, they study also compare the uses of preposition in the oral LINDSEI-NO data to the written English of Norwegians in the International Corpus of Learner English. Contrary to what they hypothesize, the proportion of inappropriate use is similar both in spoken and written discourse. In the two corpora, the inappropriate usage account for only less than 5% of the total usage. Hence, Nacey and Graedler's study questions the widespread assumption that prepositions pose a particular challenge in L2 learning. In "**L2 acquisition of temporality: Findings from a corpus based study of the grammatical encoding of past time**", *Ann-Kristin Helland Gujord* investigates the role of verb semantics (the Aspect Hypothesis) and L1 influence in texts written by Vietnamese and by Somali learners of Norwegian, extracted from ASK Norwegian learner corpus at the University of Bergen. Gujord's study first provides additional empirical evidence for the existence of particular L1-specific patterns influencing the production of L2 Norwegian temporal morphology. Perhaps more importantly, however, her study fails to support one of the core predictions of the Aspect Hypothesis – specifically, that telic verb phrases will be inflected before atelic verb phrases. Gujord suggests that her findings indicate the importance of using different data types to test the Aspect Hypothesis than have hitherto been used.

Two articles go beyond a focus on learner language alone, by comparing patterns in learner language and the language of native speakers: In "**Learners' and native speakers' use of recurrent word-combinations across disciplines**", *Signe Oksefjell Ebeling* and *Hilde*

Hasselgård investigate the use of recurrent word-combinations in texts written within two different disciplines (linguistics and business) by two groups of novice writers: Learners of English, extracted from the Varieties of English for Specific Purposes database (VESPA-NO), and native speakers of English extracted from the British Academic Written English (BAWE) corpus. The study builds on previous research documenting that n-grams may differ across disciplines, and between learners and native speakers. The picture emerging from the analyses is complex; yet, some differences are observed in the distribution of n-grams between the linguistics discipline and in the business discipline as well as between the two groups of writers. However, the most important finding is that the discipline comparison involves more statistically significant differences than the comparison between the learners and native speakers of English. In "**Patterns of misspellings in L2 and L1 English: a view from the ETS Spelling Corpus**", *Michael Flor, Yoko Futagi, Melissa Lopez* and *Matthew Mulholland* study patterns of misspellings based on the ETS Spelling corpus which consists of essays written as a part of an exam answer for a test (GRE or TOEFL). This corpus is compiled for the purpose of developing and evaluating new spell-checking software and is systematically annotated for spelling errors. Flor *et al.* explore the effects that a range of factors (i.e. proficiency, word length, edit distance) may have on frequency and type of misspellings. They find, for example, that the rate of misspellings decreases as proficiency increases, that learners produce more severe errors than native speakers do, and that word length and word frequency seem to affect misspellings. One important finding is that writing proficiency seems to be a more important factor than the learner/native distinction.

Whereas the earlier articles in this edition of BeLLS analyze corpus data to shed further light on various aspects of learner language, the final two articles deal with different issues related to learner corpora: Corpus annotation is the topic in "**How to annotate morphologically rich learner language. Principles, problems and solutions**", written by *Sisko Brunni, Liisa-Maria Lehto, Jarmo H. Jantunen,* and *Valtteri Airaksinen,* an article based on the fundamental premise that the usefulness of corpora depends on the information added to the language material, and how easily that information may be extracted. Specifically, Brunni *et al.* discuss challenges arising from the annotation process of the International Corpus of Learner Finnish (ICLFI). Finnish has a complex morphosyntactic structure, and Brunni *et al.* show that such languages represent a particular challenge as merely part–of–speech–tagging is not sufficient, for example because it does not provide enough information about the learners' case selection. The authors furthermore illustrate that the annotation of learner error is particular challenging when the target language is characterized by a rich morphology, and demonstrate how these challenges can be met. Brunni *et al.* also underline the importance of making descriptions of annotation processes available, and as standardized as possible. In "**Discriminating CEFR levels in Greek L2: a corpus-based study of young learners' written narratives**", *Maria Giagkou, Vicky Kantzou, Spyridoula Stamouli, and Maria Tzevelekou* demonstrate how learner corpora may contribute to the development of language-specific proficiency scales in the Common European Framework of Reference for Languages (CEFR). Although CEFR has gained a prominent role in language learning, teaching, and assessment in Europe since its 2001 inception, the linguistic scales suggested in the document are general rather than language-specific. Giagkou *et al.* investigate a corpus of 150 narrative texts written by young L2 learners of Greek to identify features criterial for proficiency levels A2, B1 and B2 in Greek. They find that a range of linguistic features systematically contribute to differentiate the levels – for instance, degree of subordination, use of connectives, and the frequency of correct use of clitics. Their analysis

also indicates, however, that it is more challenging to differentiate adjacent levels; that is, fewer features discriminate A2 texts from B1 texts than discriminate A2 texts from B2 texts.

Many thanks are due to the chairs of the LCR2013 program committee, Kari Tenfjord (University of Bergen), Anne Golden (University of Oslo), Fanny Meunier (University of Louvain) and Koenraad De Smedt (University of Bergen), as well as to members of the local organization committee including Victoria Rosén (chair), Gyri Smørdal Losnegaard, Karen Margrete Dregelid, Marta Olga Janik, and Marte Nordanger. We would also like to thank the many contributing authors for participating in this publication, regardless of whether their article was accepted in the end. Further thanks must be extended to our anonymous peer reviewers, who took the time to offer critical feedback on the submitted articles. The learner corpus research community is a generous one, willing and eager to donate their time and expertise when asked to do so.

Finally, we would like to encourage all interested readers to become members of the Learner Corpus Association (LCA; http://www.learnercorpusassociation.org/), an international organization officially launched at LCR2013 in Norway. LCA aims at promoting interdisciplinary learner corpus research by the following:

- supporting the compilation of learner corpora
- supporting the development of methods and tools to analyze learner data
- providing an active interdisciplinary forum for learner corpus scholars
- through maintaining a comprehensive website, and
- coordinating the Association's bi-annual conferences.

We hope to meet you all at LCR2015, LCR2017, LCR2019, and beyond!