Learner Corpus Research: LCR2013 Conference Proceedings

BeLLS 2015 Vol. 6, ISBN 978-82-998587-7-9, s. 133–152

How to annotate morphologically rich learner language. Principles, problems and solutions

*Sisko Brunni, Liisa-Maria Lehto, Jarmo H. Jantunen, and Valtteri Airaksinen**
University of Oulu

Abstract

This article illustrates the grammatical and error annotations of a morphologically rich learner language with the help of the International Corpus of Learner Finnish (ICLFI). It especially focuses on problems and solutions in morphological and error annotation, both of which are challenging due to the rich morphological structure of the target language. The article also introduces existing Finno-Ugric learner data and their annotation schemes, and compares those with the ones used in ICLFI annotations. Learner data variables, taxonomy, and principles in grammatical and error annotation are also discussed with the help of the ICLFI in the present article.

Keywords: learner corpus, corpus annotation, error tagging

**Principle contact:*
Sisko Brunni, University teacher
University of Oulu, Finland
Tel.: (+35) 8 294 483 472
E-mail: sisko.brunni@oulu.fi

Bergen Language and Linguistic Studies (BeLLS), May 30th 2015. © Brunni, Lehto, Jantunen and Airaksinen
DOI: 10.15845/bells.v6i0.812

1. Introduction

To improve the usability of corpora, it is often important to add various meta-information into data. Background information on texts, their producers and the context of data collection is essential, particularly with special corpora such as learner language material, because they are often the topic of comparative research, and commonly the texts used in comparison are chosen on the basis of background information. The usability of the data itself can also be improved by adding explanatory linguistic information. Codes or tags can be appended to words to signify, for example, the word class of a given word in its textual context (part-of-speech-tagging or POS tagging). This process is known as annotation. This term is also used for the end result of the process, i.e., linguistic tags which are attached with the electronic representation of the material (Leech 1997a, 2). (For more on the annotations process, see, e.g. Garside, Leech, and McEnery 1997).

This article begins with a brief introduction to corpus annotation followed by a description of the design and implementation process of learner data from the point of view of both grammatical and error annotation, with particular focus given to the Finno-Ugric learner data and the International Corpus of Learner Finnish (ICLFI) corpus. Finally, we outline some of the problems that have arisen during the annotation process and their solutions.

2. Grammatical and error annotation

2.1. Annotation in general

The usefulness of extensive digital corpora depends on how easy it is to extract information from them. Often, extracting information from the corpus requires that some information is added to it. For instance, homonymic expressions may belong to different word classes, and corpus users must add this information to the search results in order to use them. Annotated material already contains this information, which expedites and facilitates retrieving information from the corpus. Because annotation is time-consuming and expensive, it is not economical to repeat it over and over again. In addition, once the material has been annotated, the corpus becomes easier to utilise in the future. Previous encoding can facilitate adding new annotations or the corpus can be used for several different purposes (Leech 1997a, 4–6). For example, once both grammatical and error annotations have been added to the material, they improve the usability of the corpus and support each other in terms of searching for a particular phenomenon. Encoding word classes in sentences (POS tagging) can be used in lexicography, sentence analysis or word list generation (Leech 2004). On the other hand, Leech (2004) points out that the versatility of a corpus may not be directly proportional to the general annotations made to it, but sometimes annotations designed with a particular research context in mind may prove more fruitful. It must be noted, however, that in textual corpora the texts themselves are always the key, and annotations only provide additional information (Leech 1997a, 4).

For the annotations to be genuinely helpful, certain principles must be observed during the annotation process: 1) The annotated material must be saved in such a way that the raw data can be used at any time. Correspondingly, it must be possible to extract the annotations from the corpus and save them separately as necessary. 2) The annotation process must be carefully documented. The documentation must include details such as a description of the annotation system and information about the place of completion and the creator(s) of the annotations. In addition, factors affecting the quality of the annotation must be documented as well (possibility of errors, how they were checked, etc.) Furthermore, the annotation systems should be available to other corpus users to avoid them having to start their work from scratch. Because of this, the system should be based on a commonly approved and neutral analysis to allow for optimal, easy and extensive understanding and utilisation. No annotation system may be used as an absolute standard, because annotation needs may vary in terms of the purpose, size and language of the corpus. However, this does not mean that maximum unification should not be an objective (Leech 1997a, 6–7).

Corpora can be annotated on the basis of various principles. For example, pragmatic, discursive or phonetic annotations can be added to spoken language corpora. Pragmatic annotation focuses on the function of an expression in a given context; for example, the same expression can be both a command and a request. Discourse annotation focuses on details such as pronoun references, while phonetic annotation encodes details pertaining to the pronunciation, stress and intonation of expressions. Expression styles can be annotated as well. Syntactic annotation encodes the grammatical relations of words in sentence analysis. Lemmatisation of words is a key annotation level in all corpora, but particularly so in learner language corpora in which spelling mistakes or inflection errors can significantly enlarge the variation of used word forms. In this process, information about the base form (lemma) of a word will be appended to the inflected form used in the text. This streamlines the use of the corpus by allowing users to search for all different inflections at once. The text can also be annotated semantically, which means that homonymic expressions will include information on the semantic category to which they belong. This enables users to limit the search to only apply to a form or lemma used in a specific meaning (Leech 2004, see also Garside, Leech, and McEnery 1997). The error annotation added to learner language corpora in turn enables users to analyse the errors produced by learners and compare where and how learner language differs from native speakers' language use (Granger 2002, 14).

2.2. Learner Corpus Annotation

Learner language corpora have been compiled from students from different language backgrounds and they represent different target languages. These corpora differ from each other in terms of, for example, the amount of data processing, i.e., whether the material has been annotated grammatically or do they contain error tagging.

Tagging errors has become a key component of learner language analysis known as *computer aided error analysis* (Dagneaux, Dennes, and Granger 1998, 163). The error analysis of learner language corpora has been justified with, for example, the argument that analysing learners' errors is one of the most efficient methods for describing the characteristics and development of interlanguage. This information can then be utilised in language teaching and second language acquisition research. (Izumi, Uchimoto, and Isahara 2005, 71; Granger 2002, 14.) From the psycholinguistic viewpoint, the errors are not merely deficiencies in language skills but rather are an essential and necessary part of language development. Based on this observation, it is important to examine the learner's language system as its own system, as the interlanguage (Selinker 1972), which has its own typical features (Ellis 1990). Tagging errors has many benefits, most of which involve retrieving errors from the corpus. A fully error-tagged corpus reveals atypical forms and enables searching for errors efficiently based on error type or a specific learner group. Error-tagged corpora allow researchers to point out details such as the most frequent errors made by a group of learners and how the number and the nature of the errors alter following the development of language skills. Both predictable and fully unexpected errors can be found from the corpus. In addition, encoding allows users to find so-called zero instances where the learner has not used a word (e.g. articles or conjunctions) (Dagneaux, Dennes, and Granger, 1998, 172). Nevertheless, error analysis has been criticised as well. It has been characterized as an unscientific and confusing approach that focuses on the negative aspects of learner language (Granger 2003, 466; 2007, 54). According to Granger (2003, 466), however, errors are an integral aspect of learner language and therefore worth analysing as any other feature.

At the moment, there are several learner English corpora, of which the International Corpus of Learner English (ICLE) and the Corpus of Japanese Learner English (NICT JLE) are at least partly error annotated (see, e.g. Tono 2003, 802–803; Diaz-Negrillo and Fernandes-Dominguez 2006, 87; ICLE, Granger, Dagneaux, and Meunier 2002; NICT JLE Izumi, Uchimoto, and Isahara 2005) Learner German (e.g. FALKO, Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache) and French (FRIDA, the French Interlanguage Database) have been error annotated as well (see, e.g. Diaz-Negrillo and Fernandes-Dominguez 2006, 87). A comprehensive list of learner language corpora can be found at http://www.uclouvain.be/en-cecl-lcworld.html.

However, learner language annotation cannot be limited to error tagging. Since the phases of language learning and development are key topics for SLA research (see Ellis 1994, 73–76; Pieneman 1998), the description of the phases in terms of the relationships of various linguistic phenomena is crucial, and learner language corpora provide excellent data for studying these issues. Besides error tagging, the corpora must include various grammatical annotations in order to be used in studies focusing on language development and differences between proficiency levels. The most common linguistic annotation added to learner language corpora is POS tagging (see Rooy and Schäfer 2003; Granger 2002; Schmidt 1994). According to Granger (2002, 17), tagging word classes in learner language material clearly increases the value and comparability of corpora. Furthermore, a thorough analysis of language production also requires that sentence syntax is decoded. Sentence constituents and their relationships are a prerequisite for using corpora for such purposes as machine translation and speech recognition (Leech and Eyes 1997, 34). Similarly, syntactic encoding may benefit learner language research by extending the scope of research to the interrelationships of linguistic phenomena. However, most syntactic annotation software are mainly based on English, which means that, as such, they are unfeasible for analysing the syntax of languages with more complex morphology (Leech and Eyes 1997, 47). One notable exception, however, is the Helsinki Constraint Grammar-parser (Karlsson et al. 1995) which has been used for the syntactic annotation of languages with a morphology more complex than English (Leech and Eyes 1997, 47). In addition, word class classifications and related encoding software are usually developed on the basis of or for native languages, which causes problems in terms of annotating learner language (see Diaz-Negrillo et al. 2010; Rastelli 2009). According to Rastelli (2009), features such as too strict target language-based word class encoding are unsuitable for SLA research, because SLA research is interested in language produced by learners, and both correct and incorrect expressions are essential components of learner language. This applies to other learner language phenomena as well. So far, no fully automatic system can handle the complexity of language without making errors (Heikkinen, Lounela, and Voutilainen 2012, 373–374; Leech 1997a, 2). In existing syntactic annotation systems the respective shares of automated and manual work vary greatly, but they always involve at least a manual check-up process of the annotation (see presentation in Bateman, Forrest, and Willis 1997, 167). In POS encoding, too, a key question is how much manual work is required to edit automatic encoding (Leech 1997b, 20).

2.3. A rich morphology makes a difference: Finno-Ugric learner corpora

Morphosyntactically complex languages, such as Finnish, require an approach all of their own. POS or sentence constituent encoding alone does not provide sufficient linguistic information for researchers, because morphosyntactic case selection is often what causes problems for learners. Any annotation should therefore include even more linguistic information than provided by simple word class or sentence constituent encoding in order for researchers to focus on the desired linguistic phenomenon (see Ragheb and Dickinson 2012). Several compilers and researchers of Finno-Ugric language data have struggled with the problems that the encoding of, for example, the often opaque morphological (erroneous) forms cause. There have been several improvements, however. This section introduces Finno-Ugric learner corpora that have been useful and examples in the designing of the ICLFI tagging system, which is then described in Section 3.

The 3.3 million token Estonian Interlanguage Corpus (EIC, Eslon and Metslang 2007) at Tallinn University includes written texts in several subcorpora. It contains, for example, a large subcorpus of texts produced by Russian learners of Estonian as well as reference corpus in Russian. EIC also contains various metadata about text producers and texts as well as statistics. It is possible to view data as raw texts without tags or as syntactically and morphologically annotated texts. The data is also error annotated based on multilevel linguistic error taxonomy and a special concordance program designed for annotation is created. Marked errors can be observed in narrow contexts or in the full texts. Errors are marked in the texts, and error classes are showed in a pull down menu. (Eslon 2007, 101, 104–105; 2014, 438–439, 442; Eslon and Metslang 2007, 106–107.) The Hungarian Learner Corpus (Dickinson and Ledbetter 2012) consists of student journals from three different proficiency levels written at Indiana University. Currently, there are data from fourteen journals included, but more data is being collected. In this corpus annotation is only conducted out for error annotation.

The Corpus of Advanced Learner Finnish (LAS2, Ivaska and Siitonen 2009) was compiled at the University of Turku and consists of written academic texts of non-native speakers of Finnish. The proficiency levels of the writers are high intermediate or advanced. LAS2 also includes reference corpora of native speaker of Finnish. It allows for longitudinal research, because some of the data was collected from the same informants during a period of one to four years. LAS2 is partly annotated: data is lemmatised and annotated grammatically in terms of parts-of-speech, morphological forms and syntactic functions. The data is not error annotated, although there is room for optional comment annotation where the annotator can add error information (Ivaska 2014, 21–28). The four other existing learner Finnish corpora — the Cefling Corpus, Topling Corpus, the Finnish National Certificates learner corpus (YKI) and the Dialuki data — are not error annotated so far. All of these corpora are found at the University of Jyväskylä, and are compiled for projects where aims have included the analysis of school children's and adults' language learning and the study of the development of second language proficiency (Cefling and YKI data, see e.g. Toivola and Tossavainen 2011 and Martin et al. 2010, see also SLATE), the comparison of cross-sectional and longitudinal sequences of the acquisition of writing skills of school children and students (Topling, see e.g. Toropainen, Härmälä, and Lahtinen 2012), and the investigation of Russian speaking children's development in writing skills (Dialuki, see e.g. Nieminen et al. 2011). All of these corpora contain written data, and YKI also includes spoken language. The Cefling data contain grammatical annotation, while the other three contain raw text data without any annotation. The Cefling and Dialuki data also include comparable native language texts.

The error categorisations of Finnish and cognate languages in EIC and the Hungarian Learner Corpus are more comparable to ICLFI than those based on English due to the similar, rich morphology of the languages in question. The error annotation systems and classifications of EIC, the Hungarian Learner Corpus and LAS2 clearly differ from one another, but they have been useful when planning error annotation of the ICLFI.

In the next sections we describe the International Corpus of Learner Finnish (Section 3) followed by a description of the design and implementation process of both the grammatical annotation (Section 3.1) and the error annotation (Section 3.2). Finally, we outline some of the problems that have arisen during the annotation process together with their solutions (Section 3.3).

3. International Corpus of Learner Finnish – ICLFI

The International Corpus of Learner Finnish (ICLFI), which has been under compilation at the University of Oulu since 2007, is one of six major digital textual corpora of Finnish as a second or foreign language being currently compiled in Finland. ICLFI focuses on Finnish as a foreign language, as the texts included therein originate from students studying Finnish as a major or minor subject or learning Finnish in individual courses at tertiary level outside Finland. Table 1 presents the current status of the corpus in various figures and features.

Table 1. ICLFI in figures and features (as of September 2014)

Size - tokens - texts	 Approx.1 million tokens Approx. 6,000
Annotation - grammatical annotation - error annotation	 100% 5%
Proficiency assessment (CEFR) - A1 - A2 - B1 - B2 - C1 - C2	 0.1% 7.3% 43.2% 36.1% 11.9% 2.0%
Lemmatisation	100%
Native languages	22
Data collection	Both handwritten texts and texts composed with word processing software
Genres	Fiction and non-fiction
Assignment type	Exercise or examination completed in connection with teaching

In order to utilise the corpus optimally in a wide range of ways, particular attention must be paid to systematic data collection and the documentation of background factors. ICLFI contains ample metadata on the text producers, the context of the text production and the texts themselves. The variables have been documented as follows:

Learner variables

- Personal information: Age, place of birth, gender, place of residence
- Language proficiency: Mother tongue and other language proficiencies
- Proficiency level: According to length of studies

Learning context variables

- Exposure to the language being studied: Parents' native language(s), the use of Finnish as the language spoken at home, teaching provided by relatives (if any), residency in Finland (if any), teacher’s native language
- Textbooks used

Text variables

- Genre and topic of written assignment
- Time allocation: Limited or unlimited
- Writing context: Exercise or examination
- Use of learning aids: Dictionaries, etc.
- Place of completion: At home, in class, other location
- Proficiency level: In accordance with CEFR

Other information

- Time and place of collection
- Medium: Handwritten or produced using word processing software.

The background variables considered most extensively in research are the students' mother tongue and proficiency level. At the moment, there are texts from 22 mother tongue groups of which eight (Estonian, Russian, German, Polish, Swedish, Chinese, Czech, Dutch) form a subcorpus large enough to enable research into topics such as transfer; the number of texts in subcorpora based on other mother tongues is currently too small to allow for such research without supplementary material. However, they can be included in research requiring a large amount of learner language where mother tongue is irrelevant, or research that requires one mother tongue subcorpus and extensive general learner language material (see, e.g. Jantunen and Brunni 2012). As shown in Table 1, the majority of the material falls into proficiency levels A2-C1 of the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001). This is due to the nature of the texts; most of the types of texts included in the corpora cannot be produced by absolute beginners. It should be noted, however, that the proficiency level assigned on the basis of the CEFR describes the level of the text, not the student: each text has been assessed at least by two assessors, and different texts by the same author may be assigned a different level. If a text has obtained different gradings, a third assessment was conducted (the proportion of texts assessed three times is 3%). In addition to the CEFR level estimate, the metadata includes the amount of teaching received by the student, which can be taken into consideration when describing their proficiency level. All aforementioned metadata is listed in the metadata section of each text file before the written assignment proper.

Corpora are often described with various classification features (for corpora in general see e.g. Atkins, Clear, and Ostler [1992]; Laviosa-Braithwaite [1996] for translational corpora; Granger [2007] for learner data; see also Lehtinen, Karvonen, and Rahikainen [1995] for Finnish data). Learner language corpora can also be classified according to various dimensions; one extensive dimensional classification of learner language material can be found in the description by Jantunen (2011). According to this classification, ICLFI can be described as follows:

- genre: multi-genre
- theme: general
- register: written language
- language: monolingual
- comparability/variant: non-comparable (no native variant)
- translation: non-translational
- time: synchronic (partly diachronic)
- sample: whole-text
- medium: electronic and hand-written texts
- annotation: both annotated and raw text
- mother tongue: multi-L1
- proficiency level: multi- proficiency-level (A1–C2)
- learning context: foreign language
- learning method: more learning than acquisitional

3.1. Morphological annotation process of the ICLFI

The ICLFI utilises morphosyntactic grammatical annotation, which includes lemmatisation and the encoding of word classes, inflections, and sentence constituents. The morphological annotation is a multi-phase process in which the raw text is lemmatised and grammatical information is added to it through encoding. Lemmatising is especially important in the corpus because of the complexity of Finnish declension and conjugation systems which can change the roots of words so crucially that searching for some particular word from the corpus might later become too complicated. Similar to grammatical annotation in general, this process is partly automated, although the data processing also includes a manual check-up phase. The morphological annotation of learner Finnish is challenging because an automatic computer-generated analysis does not yield as good results about learner-produced material as it would about material produced by native speakers. In addition, the rich morphology of the Finnish language, which poses a great challenge to learners, leads to erroneous combinations of lexical and grammatical morphemes, which then are misinterpreted by the automatic analyser. Fully manual annotation would be far too laborious, so the ICLFI corpus has been encoded semi-automatically. For further information about the annotation process, see De Haan (2000, 71), Jelínek et al. (1999, 132–133) and Leech (1997a, 8).

Particularly problematic items in learner language annotation are erroneous forms produced by learners. With the ICLFI corpus this problem has been solved by exporting the text file to Microsoft Word before the automatic encoding process. This is when spelling mistakes and problems with inflected forms are removed from the text. Microsoft Word helps in this process by automatically underlining the mistakes in red. The errors found by Microsoft Word, such as quantity and gradation errors, are placed inside angle brackets before the target form (see examples 1 and 2), which causes the syntactic parser to ignore them during the automatic annotation process. The purpose of this phase is to edit the text to an extent which is enough for the annotation software to read and analyse it inasmuch as the software wouldn't be able to understand the incorrect forms produced by learners (Jantunen 2011, 98).

(1) Minun <kodussa> kodissa monet kirjat.
'There are many books in my house' (**kodussa* is misspelled)
(2) <Sängi> Sänky on iso ja mukava.
'The bed is big and comfortable' (**sängi* is misspelled)

The aim is not to correct errors by changing the word or inflect it to suit the context. Even in cases where the inflection did not fit the context and Microsoft Word underlined the problematic sections in green, the errors were not corrected for the annotation software (see example 3). In other words, any errors are corrected as little as possible.

(3)	Menen	
	ostamaan	
	valkosipul*iin*	@NH N SG ILL (*illative*, pro *genitive*)

*'I'll go to buy to garlic (*valkosipuliin*)' pro 'I'll go to buy garlic (*valkosipulin*)'

Once the errors have been corrected temporarily, texts have been annotated using a parser application. Annotation produces metainformation including lemmas and morphosyntactic encoding. After the automatic encoding the errors must be restored to the text files in their original formats because the original incorrect forms inhere in the corpus.

Since the automatic encoding process always results in some errors, the final phase of the process is the manual check of the morphosyntactic encoding. This is the most time-consuming phase of the grammatical annotation process. The parser may provide many alternative encodings for a single form, and the annotator must select the correct alternative manually. Annotators can also add alternative morphological interpretations to problematic expressions. In example 4 below, for instance, the possible interpretations for the verb (*katsoan*) include 1st infinitive form and personal suffix or 1st person singular (for a more detailed description of annotation codes and definitions, see Appendix 1). When reviewing the analysis, the annotator can also add several alternative lemmas and morphological interpretations (see example 5), which researchers can later use as a basis for various searches (Lehto, Brunni, and Jantunen 2013).

(4)	Minä	@NH PRON SG P1 NOM
	katsoan	@MAIN V ACT INF F1 SG P1
		@MAIN V ACT IND PRES SG P1
	televisiota	@NH N SG PTV

'I watch television'

(5)	tulevana		@PREMOD N SG ESS
	vuonna		@NH N SG ESS
	touhikuussa	*touhikuussa	@HEUR
		toukokuu	@NH N SG INE

'in the coming year in ?/May'

The end result of the checking process is a lemmatised and grammatically encoded version of the original text. The lemma is the option that is most readily visible from the text, rather than the one that best fits the context (see example 6).

(6)	Kotini	koti	@NH N SG NOM CLI POSS P1
	sijoittaa	sijoittaa (pro *sijaitsee*)	@MAIN V ACT IND PRES SG P3
	Tartossa	Tartto	@NH N SG INE PROP

*'My home locates (*sijoittaa*) in Tartu' pro 'My home is located (*sijaitsee*) in Tartu'

Minor inflection or spelling errors (such as gradation and quantity errors) do not change the lemma, if the lemma is obvious from the context (see example 7).

(7)	Jouluna	joulu	@NH N SG ESS
	me	me	@NH PRON PL P1 NOM
	onneksi	onneksi	@ADVL ADV
	tapamme	tavata (not *tappaa*)	@MAIN V ACT IND PRES PL P1
	kaikki	kaikki	@NH PRON NOM

*‘At Christmas we’ll fortunately kill (*tapamme*) everyone’ Should be: ‘At Christmas we’ll fortunately meet (*tapaamme*) everyone’

Unidentified words are marked with a HEUR tag. This is used for adding information about foreign-language words as well (see example 8).

(8)	On	@MAIN V ACT IND PRES SG P3
	muodostunut	@MAIN V ACT PCP PAST
	kielibarjääri	@NH HEUR N SG NOM

‘A language barrier is formed’ (**kielibarjääri* is not a Finnish word)

The software does not necessarily recognise the titles of books, television series or films as proper nouns, so these are marked manually with a PROP tag that signifies proper names. Greetings and interjections (*hei* ‘hello’, *huomenta* ‘good morning’) are tagged as interjections with the INTERJ tag. There is no separate tag for colloquialisms.

The syntactic parser makes some recurring errors. For example, it interprets the *minä* (‘I’) pronoun at the beginning of a sentence as the essive form of the *mikä* (‘what’) pronoun. The parser often interprets homonymic expressions erroneously or provides two separate interpretations. Words at the beginning of a sentence are sometimes analysed as proper names. Because the parser offers different options for the annotator to choose from and also has a tendency to repeat errors, some decisions have been made to simplify the annotation process.

All the syntactic parsers are created on the basis of a grammar. The one behind the parser used in ICLFI does not completely follow the one employed in the annotation correction process, and that causes some systematic corrections like marking the modifiers. The morphosyntactic annotation process of ICLFI corpus strives to follow the classification presented in the grammar *Iso suomen kielioppi* (The Comprehensive Finnish Grammar, Hakulinen et al. 2004). However, some exceptions have been made. For example, there is no separate tag for particles (with the exception of interjections), and ordinal numbers are considered numerals rather than adjectives. Prepositional or postpositional complements are also tagged as heads to facilitate the encoding. Because the issue at stake is particularly one of learner language, some extra information with tags or second options has been added to help researchers. The general principles, recurring errors that need to be disambiguated, and the annotation scheme for the grammatical annotation and any deviations from the classification presented in *Iso suomen kielioppi* have been documented in the (as yet unpublished) annotation manual of the ICLFI project.

3.2. Error annotation systems in FU (Finno-Ugric) learner data and error classification of the ICLFI

In the Hungarian Learner Corpus, annotation is carried out using EXMARaLDA (Extensible Markup Language for Discourse Annotation), which allows for multiple simultaneous tiers of annotation. In the annotation scheme of the learner Hungarian, corpus annotation categories are distinguished from annotation layers. Firstly, there is the error layer, which includes different error categories such as morphological errors, and secondly there is the adjustment layer. The adjustment layer enables making alterations necessitated by correction of a linked error (Dickinson and Ledbetter 2012, 1660–1661).

The Estonian Interlanguage Corpus (EIC) is partly error tagged. It has its own concordance program specifically designed for it and it allows finding errors according to error classes. Marked errors can be observed in narrow contexts or in full texts. Errors are marked in the texts, but error classes are shown in a pull down menu. (Eslon 2007, 101, 104–105; 2014, 442.) The Corpus of Advanced Finnish Learner is not currently error annotated, but there is an optional comment annotation in which the annotator can add error information, so error annotation could be done in the future (Ivaska 2014, 27).

The development of the error annotation system for ICLFI started in early 2013. In the course of a year, we have created a functional classification and error encoding system. As of September 2014, there are some four hundred error annotated texts, with a total of 48,000 tokens. This comprises approximately five percent of the total number of tokens in ICLFI (see Table 1). The error annotated texts were written by Swedish, Dutch and Estonian learners. At the moment, errors are encoded into the ICLFI manually; they are tagged directly into the text file and there are no tools used specifically for tagging and correcting errors (cf. the Louvain error editor, Granger 2002, 19–20).

The error annotation tags and classifications of the ICLFI were designed with the help of previous classifications. According to Eslon and Metslang (2007, 107), error classification that is divided into error categories and finer subcategories makes it possible to illustrate the multidimensionality of errors. The error classification system used in ICLFI is based on error type, i.e., whether the errors are lexical or syntactic (for more information about error types, see Granger 2002, 19). The rich and diverse morphology of the Finnish language has been taken into consideration in the design and development process of the classification system. Both the Estonian Interlanguage Corpus and the Hungarian Leaner Corpus use error classification which consists of several classes and subclasses. Some of these are more language-specific error classes, such as vowel harmony, and some are more universal, such as agreement errors. (Dickinson and Ledbetter 2012, 1660–1661; Eslon 2007, 101–102.) The classification system of ICLFI is also hierarchical, covering all levels of language from phonology to syntax, lexis and phraseology. For example, morphosyntactic errors form one main error category, under which fall such issues as the number and case of objects (for further information on the error categories of other corpora, see, for example, Granger 2003, 467).

After a preliminary review of the existing FU learner data error annotation systems, several error schemes for ICLFI were designed and tested using a small amount of text material. The classifications and the encoding system itself were then outlined based on these experiments. The current error classification system is presented in Table 2.

Table 2. Error classifications in ICLFI

1 ORTHOGRAPHIC	1A spelling 1B punctuation 1C compounding
2 PHONOLOGICAL	2A quantity 2B diphthong
3 MORPHOPHONOLOGICAL	3A consonant gradation 3B vowel harmony
4 MORPHOLOGICAL	4A nominal inflection, form 4B nominal inflection, use 4C verbal inflection, form 4D verbal inflection, use 4E indeclinable, form 4F indeclinable, use
5 MORPHOSYNTACTIC	5A possessive suffix 5B congruence 5C subject case and number 5D object case and number 5E predicative case and number 5F adverbial case and number 5G case government (rection)
6. SYNTACTIC	6A word order 6B non-finite forms and clauses 6C phrase 6D sentence type 6E unnecessary word
7 LEXICAL	7A noun formation 7B verb formation 7C word choice 7D word coinage 7E style and register 7F foreign word 7G missing word
8 PHRASEOLOGICAL	8A phraseology
9 UNEXPLAINABLE	9A unexplainable

Error categories have been criticised for being insufficiently defined, subjective and based on mixed criteria (Dagneaux, Dennes, and Granger 1998, 164). One of the key elements of a functional error annotation system is uniformity; the detailed descriptions of errors, the definitions of different error categories, and the encoding principles should be outlined in the error annotation manual. This is one method of minimising subjectivity (Granger 2003, 467, see also Dickinson and Ledbetter 2012, 1660). The subjective interpretation of the annotator has been considered in the ICLFI error annotation system, and the reliability of the material has been improved by compiling an error annotation manual, as well as by deciding on error annotation solutions and the compilation and editing of the corpus in ICLFI project meetings. The ICLFI error annotation manual was compiled after some of the annotation was already completed. This provided us with a description of the contents of the error categories and allowed us to specify and improve the categorisation. The error manual contains a description and examples of errors belonging to each category.

One aim of the annotation was to make the error tags used in ICLFI universal; according to Granger (2003, 467), error categories should be reusable and general enough to be used for several different languages. However, learner language corpora have previously been limited to certain specific languages, and there has been little error annotation of morphologically varied languages (Dickinson and Ledbetter 2012, 1659). Due to the rich morphology of the Finnish language, systems created for Indo-European languages (see, e.g. ICLE, Granger, Dagneaux, and Meunier 2002) are not directly applicable as a basis of the error categorisation used in ICLFI. This is due to the fact that the errors and error patterns of morphologically varied languages are different compared to, for example, fusional languages (Dickinson and Ledbetter 2012, 1659). For instance, when inflecting words and combining morphemes, learners may have produced ambiguous or opaque erroneous forms. The errors in ICLFI can be roughly divided into three categories in terms of the difficulty of the annotation:

1. Unambiguous errors: clear and easily classified cases
Errors in vowel harmony or possessive suffixes belong to this category (example 9). These errors can also be morphosyntactic (example 10):

(9) Ensimmäisessä kerrokse*ssä* <err=F 'kerrokse*ssa*'_MF_VH> 'On the first floor' → vowel harmony error.

(10) Minun huonee*lla* <err=U 'huonee*ssani*'_MSYN_ADVLI_INE+MSYN_REF_POSS_P1> on yksi ikkuna.'There is one window in my room.' → morphosyntactic error: incorrect case in adverbial and missing possessive suffix.

*Minun *huonee-lla* on yksi ikkuna.
My *room-ADE* is.3SG one window.

Minun *huonee-ssa-ni* on yksi ikkuna.
room-INE-POSS.1SG

2. Ambiguous errors: cases with several alternative interpretations
In such cases, it is difficult to discern which category the error belongs to. Alternative interpretations are coded and the researcher has to decide whether it is, for example, a quantity error or an inflection error in question:

(11) Ensin *pannan* <err=F'*pannaan*'_PHON_QV\F'*panen*'_MORF_V_INFL_SG_P1> kahvi tulelle.

*Ensin *pan-na-n* kahvi tule-lle.
At first *put-INF-1SG* coffee.NOM fire-ALL.

Quantity error?	Inflection error?
Ensin pan-n*a*-*a*n kahvi tulelle.	Ensin pan*e*-*n* kahvin tulelle.
put-PASS-PASS	*put- 1SG*
'At first coffee (pot) *is put* on fire.'	'At first *I put* coffee (pot) on fire.'

3. Undefined errors: cases where the error type cannot be categorized
The sentences in question are often so unclear that the errors cannot be categorized. Often the context is not helpful either. These errors can be easily over-interpreted based on the expected correct form:

(12) *Se on ei hyvää.* 'It is no good**'** → Is there an error in the predicate or is the meaning 'non-good', in which case it is a compounding error?

*Se on *ei hyvä-ä*
It is.3SG *no good-PTV.*

Se *ei ole hyvä-ä.*	Se on *ei-hyvä-ä.*
It *NEG is.3SG good-PTV.*	it is.3SG *non-good-PART.*
'It is not good.'	'It is non-good.'

As shown in the examples 11 and 12, the error annotation system used in ICLFI enables researchers to consider overlapping errors. Milton and Chowdhury (1994, 129) have written about the uncertainties pertaining to error categorization, because it is not always possible to place errors into a single category. According to them, the encoding system should enable adding several possible interpretations (see also Dickinson and Ledbetter 2012, 1662 and Granger 2003, 467).

In the ICLFI error classification, a single error may belong to several categories depending on the linguistic level on which it is examined (particularly errors belonging to the 2nd category for ambiguity). The error encoding takes this into consideration by providing annotation options. In unclear cases (undefined errors) the researcher must ultimately decide the category of the error in question. Some of the alternative interpretations may seem pointless, but in reality the annotator sometimes cannot discern which mistake the learner has made, not even based on the form in the text. As Milton and Chowdhury (1994, 129) point out, despite attempting to add all key interpretations, the analysis is unlikely to ever cover all possible alternatives.

In practice, errors are annotated by adding an error tag after the morphological annotation in the text file. The error tag contains information about the target form: a correction (if possible), as well as information of the error category in question and a morphological description of the desired form. Line P4 in example 13 presents an example of an error tag.

(13) <P1>Minä<bf=minä> <@subj_PRON_SG_P1_NO
<P2>en<bf=ei> <@pred_Aux_V_ACT_SG_P1>
<P3>tarvitse<bf=tarvita> <@V_ACT_PRES_NEG>
<P4>*kengät*<bf=kenkä> <@obj_N_*PL_NOM*> **<err=U'*kenkiä*'_MSYN_OBJ_PL_PTV>**
'I do not need shoes.'

First it shows whether the error pertains to form (F) or use (U). Next, it shows the target (correct) form *kenkiä* ('shoes') (because the object of a negative sentence must be in the partitive case). The section MSYN_OBJ shows the error category: in this case, the main category is morphosyntax, under which belong object case and/or number errors. Finally, the tag shows that the target form added by the annotator is the plural partitive. The grammatical and error annotation in ICLFI provide versatile search options: it is possible to conduct searches using morphological or error tags either separately or together with the search term. Example 13 shows a case where it is possible to search for, for example, object errors by combining information about the form produced by the student (grammatical annotation, PL_NOM) and about the desired form (error tag, PL_PTV).

3.3. Problems, solutions and principles of error annotation

According to Granger (2003, 467), error annotation should be informative, i.e. detailed enough to provide information about errors made by the learners. At the same time, its information content should be succinct enough to make the error annotation system easy for the annotator. The error annotation system used in ICLFI aims to provide sufficient options for researchers but eliminate unnecessary or irrelevant interpretations. After examining the problems pertaining to error annotations and considering possible solutions, some key error annotation principles have emerged: the principles of context, simplicity, the avoidance of error accumulation and exactitude - similar to those Dickinson and Ledbetter (2012, 1662) presented earlier when error annotating Hungarian language and finding solutions for multiple analysis.

The context principle refers to the use of context to facilitate interpretation: errors are encoded based on their most likely interpretation. The context-based interpretation takes priority, and over-interpretation and guessing the author's intention are avoided, whenever possible. The simplicity principle calls for annotators to strive towards finding the easiest interpretation of a given error. If the category of an error is easily determined, overly complex interpretations should be avoided. The error annotation used in ICLFI takes into consideration only as many errors as necessary; in other words, the accumulation of errors made by the learners is avoided. For example, if a modifier is in congruence with its head, but the case of the head is false, only the head is tagged as erroneous and the modifier is left untagged. Dickinson and Ledbetter (2012, 1662) also prefer fewer corrections to benefit the learner. The annotation program they use allows for the possibility of adding adjustments to the annotation. As a consequence, they refer to correction linked to previously-annotated errors as adjustments rather than error per se.

The exactitude principle in ICLFI error annotation refers to the fact that some error categories provide more information about the error type than others. For example, noun formation errors are a fuzzier category than diphthong errors. That is because any problem in the stem of a noun could be classified as noun formation error whereas diphthong error only includes problems in certain vowel combinations. Also the category of spelling is intended for a limited type of errors. According to Dickinson and Ledbetter (2012, 1661–1662), in the Hungarian Learner Corpus, for example, phonology errors could also be considered as spelling errors. However, many phonological features such as vowel length are contrastive in the Hungarian language whereas spelling errors do not generally cause a change in meaning. They are therefore classified as phonology errors, not spelling errors. In the same way, the error annotation code for spelling errors is not used in ICLFI if there is a more suitable error class to be found (see example 11). However, the exactitude principle does not negate the fact that errors may overlap, in which case no single interpretation is better than another.

4. Conclusion

The description of the annotation process presented here shows that annotation is a challenging task. The annotation of learner language material poses more problems than the annotation of so-called native language material, particularly because forms produced by language learners often differ greatly from the target language forms. Since grammatical annotation is rarely sufficient for describing learner language material, an accompanying error annotation must be completed as well. This increases the annotation of the material. Despite its arduousness, grammatical and error annotation should be completed for learner language data, because it improves the usability of the material considerably: The material can be studied from several different angles and with several different methods, which may yield both qualitatively and quantitatively versatile information about the learner language.

Since learner data annotation is a time-consuming, laborious and complex process, and the work done with one set of data may benefit other corpus compilers. This is only possible if the schemes embodied in annotations are documented and available. The annotation manual helps researchers to interpret the annotation and the query results, and it also helps compilers and annotators of other learner data to start and plan their annotation. Thus, work done with one set of data provides a platform of annotation upon which further annotations can build. However, since it remains the case at the moment that annotation procedures often vary noticeably from data to data, explicit documentation is needed. To make the existing learner data more comparable, it would, of course, be useful for the various annotation systems to be more or less similar and uniform. A detailed and explicitly described manual will help harmonize the annotation schemes. This does not apply only to Finnish learner data but also to data of related languages. It is also essential that the annotation scheme be included in the meta-information when ICLFI is relocated to the Language Bank of Finland (CSC) in the near future. The ICLFI error annotation manual and principles should provide valuable information to those who are launching an error annotation process of Finnish or related learner language data.

References

Atkins, S., J. Clear, and N. Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7 (1): 1-16.

Bateman, J., J. Forrest, and T. Willis. 1997. The use of syntactic annotation tools: Partial and full parsing. In *Corpus annotation. Linguistic information from computer text corpora*, eds. R. Garside, G. Leech, and A. McEnery, 1-18. New York: Longman.

CEFLING = *Linguistic Basis of the Common European Framework for L2 English and L2 Finnish*.(https://www.jyu.fi/hum/laitokset/kielet/tutkimus/hankkeet/paattyneet-hankkeet/cefling/en)

CSC = The Language Bank of Finland. IT Center for Science. (https://www.csc.fi/-/kielipank-1)

Council of Europe 2001. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press

Dagneaux, E., S. Dennes, and S. Granger. 1998. Computer-aided error analysis. *System* 26 (2): 163-174.

Dickinson, M., and S. Ledbetter. 2012. Annotating errors in Hungarian learner corpus. *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*. Stroudsburg: Association for Computational Linguistics, 1659-1664.

de Haan, P. 2000. Tagging non-native English with the TOSCA-ICLE tagger. In *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20)*, eds. C. Mair, and M. Hundt, 69-79. Amsterdam: Rodopi.

Díaz-Negrillo, A., and J. Fernandes-Dominguez. 2006. Error tagging systems for learner corpora. *Resla* 19:83-102.

Díaz-Negrillo, A., D. Meurers, S. Valer, and H. Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* 36 (1-2): 139-154. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair, edited by María Moreno Jaén and Carmen Pérez Basanta). (http://www.sfs.uni-tuebingen.de/~dm/papers/diaz-negrillo-et-al-09.html)

Ellis, R. 1990. *Instructed second language acquisition*. Oxford: Basil Blackwell.

Ellis, R. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.

Eslon, P. 2007. Õppijakeelekorpused ja keeleõpe [Learner corpora and language learning]. In *Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine.[Optimality, design and use of the language corpora of the University of Tallinn]*,ed. P. Eslon, 87-120. Tallinn: Tallinna Ülikooli Kirjastus.

Eslon, P. 2014. Estonian Interlanguage Corpus. *Language and Literature* 6: 436-451.

Eslon, P., and H. Metslang. 2007. Learner language and Estonian Interlanguage Corpus. In *Eesti rakenduslingvistiika ühingu aastaraamat 3 - Estonian Papers in Applied Linguistics 3*, eds. H. Metslang, M. Langemets, and M-M. Sepper, 99-116. Tallinn: Eesti Keele Sihtasutus.

Garside, R., G. Leech, and A. McEnery, eds. 1997. *Corpus annotation. Linguistic information from computer text corpora.* New York: Longman.

Granger, S. 2002. A Bird's-eye view of learner corpus research. In *Computer learner corpora, second language acquisition and foreign language teaching*, eds. S. Granger, J. Hung, and S. Petch-Tyson, 3-33. Amsterdam: John Benjamins.

Granger, S. 2003. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20 (3): 465-480.

Granger, S. 2007. A Bird's-eye view of learner corpus research. In *Corpus linguistics: Critical concepts in linguistics* 2., eds. W. Teubert, and R. Krishnamurthy, 44-72. London, New York: Routledge.

Granger, S., E. Dagneaux, and F. Meunier. 2002. *International Corpus of Learner English.* Version 1.1. Université catholique de Louvain: Centre for English Corpus Linguistics.

Hakulinen, A., M. Vilkuna, R. Korhonen, V. Koivisto, T-R. Heinonen, and I. Alho, eds. 2004. *Iso suomen kielioppi [The Comprehensive Finnish Grammar]*. Helsinki: Suomalaisen Kirjallisuuden Seura.(http://scripta.kotus.fi/visk/etusivu.php)

Heikkinen, V., M. Lounela, and E. Voutilainen. 2012. Automaattinen analysaattori tekstilajitutkimuksessa. [Automatic analyser in genre analysis]. In *Genreanalyysi – tekstilajitutkimuksen käsikirja. [Handbook of Genre Analysis]*, eds. V. Heikkinen, E. Voutilainen, P. Lauerma, U. Tiililä, and M. Lounela, 372-391. Kotimaisten kielten keskuksen julkaisuja 169. Helsinki: Gaudeamus.

Ivaska, I. 2014. The Corpus of Advanced Learner Finnish (LAS2): Database and toolkit to study academic learner Finnish. *Apples – Journal of Applied Language Studies* 8 (3): 21-38. (http://apples.jyu.fi/issue/view/15)

Ivaska, I., and K. Siitonen. 2009. Syntactically encoded learner language corpus: opportunities and questions. In The methodology of corpus studies and the problems of the coding. *Proceedings of TLU Institute of Estonian Language and Culture 11*, eds. P. Eslon, and K. Õim, 54-71. Tallinn: Tallinna Ülikooli.

Izumi, E., K. Uchimoto, and H. Isahara 2005. Error annotation for corpus of Japanese learner English. In *Proc. of 6th International Workshop on Linguistically Annotated Corpora*. Jeju Island: South Korea, 71-80.

Jantunen, H. J. 2011. Kansainvälisen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi [International Corpus of Learner Finnish (ICLFI): typology, variables and annotation]. In *Lähivõrdlusi. Lähivertailuja* 21, eds. A. Kaivapalu, J. Laakso, P. Muikku-Werner, and M-M. Sepper, 86-105. Tallinn: Eesti Rakenduslingvistiika Ühing.

Jantunen, J. H., and S. Brunni. 2012. Morfologinen priming ja fraseologia vieraan kielen oppimisessa: korpustutkimus oppijansuomesta [Morphological priming and phraseology in second language acquisition: A corpus-study in learner language]. In *Lähivõrdlusi. Lähivertailuja* 22, eds. A.Kaivapalu, P. Muikku-Werner, J. H. Jantunen and M-M. Sepper, 71-100. Tallinn: Eesti Rakenduslingvistiika Ühing.

Jelínek, T., B. Štindlová, A. Rosen, and J. Hana. 1999. Combining manual and automatic annotation of a learner corpus. *Proceedings of the Text, Speech and Dialogue: Second International Workshop, TSD'99 September 13.–17.* Plzen: Czech Republic, 126-134.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila, eds. 1995. *Constraint grammar: A language-independent system for parsing unrestricted text*. Berlin: Mouton de Gruyter.

Laviosa-Braithwaite, S. 1996. *English Comparable Corpus (ECC): A resource and a methodology for the empirical study of translation*. Unpublished PhD Thesis. Manchester: UMIST.

Leech, G. 1991. The State of the art in corpus linguistics. In *English corpus linguistics. Studies in honour of Jan Svartvik*. eds. K. Aijmer, and B. Altenberg, 8-29. London: Longman.

Leech, G. 1997a. Introducing corpus annotation. In *Corpus annotation. Linguistic information from computer text corpora*, eds. R. Garside, G. Leech and A. McEnery, 1-18. New York: Longman.

Leech, G. 1997b. Grammatical Tagging. In *Corpus annotation. Linguistic information from* computer text corpora, eds. R. Garside, G. Leech and A. McEnery, 20-33. New York: Longman.

Leech, G. 2004. Adding linguistic annotation. In *Developing linguistic corpora: a guide to good practice*, ed. M. Wynne, 17–29. Oxford: Oxbow Books. (http://www.ahds.ac.uk/guides/linguistic-corpora/chapter2.htm)

Leech, G., and E. Eyes. 1997. Syntactic annotations: Treebanks. In *Corpus annotation. Linguistic information from computer text corpora*, eds. R. Garside, G. Leech, and A. McEnery, 34-52. New York: Longman.

Lehtinen, M., P. Karvonen, and T. Rahikainen. 1995. *Tekstikorpukset [Text corpora]*. Helsinki: The Institute for the Languages of Finland.

Lehto, L-M., S. Brunni, and H. J. Jantunen. 2013. How to Annotate Morphologically Rich Language? Problems and Solutions. Poster presented at *Learner Corpus Research Conference 2013*. Bergen/Os, Norway.

Nieminen, L., A. Huhta, R. Ullakonoja, and J.C. Alderson. 2011. Toisella ja vieraalla kielellä lukemisen diagnosointi: Dialuki-hankkeen teoreettisia ja käytännöllisiä lähtökohtia. [Diagnosis of reading in second and foreign language: The theoretical and practical starting points of the Dialuki project.] In *AFinLA-e* 3, eds. E. Lehtinen, S. Aaltonen, M. Koskela, E. Nevasaari and M. Skog-Södersved, 102-115.

Martin, M., S. Mustonen, N. Reiman, and M. Seilonen. 2010. On becoming an independent user. In *Communicative proficiency and linguistic development, intersections between SLA and language testing research*. EUROSLA Monograph Series 1, eds. I. Bartning, M. Martin, and I. Vedder, 57-80. European Second Language Association.

Milton, J., and N. Chowdhury. 1994. Tagging the interlanguage of Chinese learners of English. In *Entering text*, eds. L. Flowerdew, and A. Tong, 127-143. Hong Kong: The Hong Kong University of science and technology.

Pieneman, M. 1998. *Language processing and second language development: Processability theory*. Amsterdam: John Benjamins.

Ragheb, M., and M. Dickinson. 2012. Defining syntax for learner language annotation. *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Poster Session*. Mumbai, India, 965-974.

Rastelli, S. 2009. Learner corpora without error tagging. *Linguistic Online 38*, 2/2009. (http://www.linguistik-online.de/38_09/rastelli.html)

van Rooy, B., and L. Schäfer. 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In *Lancaster University Centre for Computer Corpus Research on Language Technical Papers* 16: 835-844. (Proceedings of the Corpus Linguistics 2003 Conference, eds. Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery. (http://www.corpus4u.org/forum/upload/forum/2005092023174960.pdf)

Schmidt, H. 1994. Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester: UK.

Selinker, L. 1972. Interlanguage. *International Review of Applied Linguistics* 10: 209-241.

SLATE = *Second language acquisition and testing in Europe*. (http://www.slate.eu.org/index.htm)

TOPLING = *Paths in Second Language Acquisition*. (https://www.jyu.fi/hum/laitokset/kielet/tutkimus/hankkeet/topling/en)

Toivola, S., and H. Tossavainen. 2011. Opiskelijoiden käsityksiä yleisten kielitutkintojen korpuksen käyttömahdollisuuksista. [Students' perceptions of usability of the Finnish National Certificates learner corpus.] In *AFinLA-e* 3, eds. E. Lehtinen, S. Aaltonen, M. Koskela, E. Nevasaari and M. Skog-Södersved, 158-169.

Tono, Y. 2003: Learner corpora: design, development and applications. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, UK, 28-31 March, 800-809.

Toropainen, O., M. Härmälä, and S. Lahtinen. 2012. Kaksi asteikkoa, kaksi eri tilannetta: äidinkielellä ja vieraalla kielellä kirjoitettujen tekstien kriteeripohjaisen arvioinnin haasteita. [The challenges of the usage of two CEFR-based rating scales in assessing L1 and L2 texts in Swedish.] In *AFinLA-e*: 4, eds. L. Meriläinen, L. Kolehmainen and T. Nieminen, 60-79.

APPENDIX 1. Annotation tags and their definitions.

#	compound
(INF F4)/	-deverbal noun with the suffix *-minen*
@ADVL	adverbial
@CC	co-ordinating conjunction
@MAIN	verb
@NH	head
@PREMARK	preposition, postposition or conjunction
@PREMOD	modifier
<p>	end of paragraph
<s>	end of sentence
A	adjective
Abbr	abbreviation, e.g. EUR
ABE/	abessive
ABL/	ablative
ACC/	accusative
ACT/PASS	active/passive
ADE/	adessive
ADV	adverb
ALL/	allative
Aux	auxiliary verb (negative verb, the perfect and pluperfect tense of the *olla* verb)
CARD/ORD	cardinal number/ordinal number
CLI	clitic particle
CMP/SUP	comparative/superlative
COM/	comitative
CS	subordinating conjunction
ELA/	elative
ESS/	essive
GEN/	genitive
Heur	unknown word
ILL/	illative
IND/IMP/CND/SUB	indicative/imperative/conditional/potential
INE/	inessive
INF 5/	-the *-maisillaan* form
INF F1/	A infinitive
INF F2/	E infinitive
INF F3/	MA infinitive
INS	instructive
INTERJ	interjection
KAAN /-KIN/-S/-PA/ -HAN/-KO/-KA	
N	noun
NEG	the verb following the negative verb
NOM/	nominative
NOM/GEN/PTV...	case
NUM	numeral
P1/P2/P3	1st person/2nd person/3rd person
PCP AGT	agent participle
PCP PAST/	NUT participle
PCP PRES/	VA participle
POSS P1/P2/P3	1st /2nd/3rd person possessive suffix
POST	postposition
PREP	preposition
PRES/PAST	present/past tense
PRON	pronoun
Prop	proper name or noun
PTV/	partitive
SG/PL	singular/plural
TRA/	translative
V	verb