Syntactic Annotation of Learner Corpora

Victoria Rosén and Koenraad De Smedt, University of Bergen

Abstract

Syntactic annotation of learner corpora is useful for investigating the grammatical properties of learner language. We discuss two approaches to syntactic annotation based on different methodological choices. One approach, recently proposed in the literature, is the manual annotation of learner language with dependency relations. Another approach, which we present as an alternative, is based on automatic parsing of a 'correct' version with an L2 grammar.

Introduction

Learner corpora have been recognized as providing a valuable empirical basis for addressing theoretical questions as well as developing practical applications. Findings from learner corpora may promote the formulation of more informed and precise models of second language acquisition, which in turn may lead to the development of more effective teaching materials and tools.

Among the theoretical issues that could benefit from corpus research are hypotheses related to the potential transfer of linguistic analytical and productive approaches from L1 to L2, and the possible identification of stages of learning related to specific language characteristics. The investigation of such hypotheses benefits hugely from high quality annotation of learner corpora. In particular, searching for specific items or patterns in large samples of learner language can provide much more informed results as more information is added to the corpus.

The annotation of learner corpora has been focused mostly on marking so-called errors or nonstandard language use. An overview of the types of annotation available in learner corpora is provided by Granger (2008). Even if error annotation makes it easy to search and quantify divergences from the norm, a usual limitation of error annotation is that it only marks such divergences, and does not facilitate the study of the nature and extent of correct L2 use. Furthermore, error annotation normally treats divergences as locally marked phenomena, an approach which by itself does not support searches which take into account the full and detailed syntactic contexts in which the errors occur.

A more fine-grained study of all grammatical aspects of learner language, both correct and incorrect use, is desirable. For instance, in studying missing subject-verb inversion in Norwegian, it could be interesting to find out what kind of initial phrases occur in sentences where learners do not apply inversion. Furthermore, it would also be interesting to find out where learners correctly apply inversion. However, error annotation does not support the study of other properties of the learners' language which are *not* errors, in particular properties related to the extent of correct language use at various stages of learning, for instance when certain constructions are correctly used or what the complexity or fluency of the language is at certain stages.

The ASK corpus (Tenfjord et al., 2006b,a) is widely recognized for its error annotation in texts written by learners of Norwegian. It contains manual markup of errors as well as a 'correct' version, i.e. an edited and normalized version of the texts in correct Norwegian, based on a possible interpretation of the learner's language. However, as it does not contain markup beyond the level of single words, we want to address the question of how such markup can be efficiently and accurately added in order to extend the possibilities of the corpus.

In the rest of this article, we will first discuss some choices involved in the construction of treebanks, i.e. syntactically annotated corpora. We will then discuss a recent proposal in the literature based on the manual annotation of learner language with dependency relations (Dickinson and Ragheb, 2009). Then we will present an alternative approach based on automatic parsing of a 'correct' version with an L2 grammar. Both approaches are based on established treebanking methodologies, but neither has as yet been tested on significant amounts of learner language. Our paper is meant as a comparison of the theoretical soundness and the potential effectiveness of the two methods.

Treebanks

Syntactic markup of a corpus improves the ability to search and retrieve syntactic constructions. Many linguistic corpora are at present annotated at word level only. They are typically lemmatized (i.e. all word forms are marked with corresponding citation forms) and annotated with part of speech (POS) tags for each word. The annotation in ASK also includes 'shallow' markup of some grammatical functions. Shallow annotation increases the linguistic information value of the corpus over plain text, but it provides only limited assistance for finding phrasal (hierarchical) structure. Automatic shallow tagging implies the risk of substantial errors, in particular for infrequent but interesting constructions, and even more so when tagging incorrect language use. Furthermore, the word level annotation in ASK does not provide reliable 'deep' syntactic or semantic information, for instance all grammatical functions, predicate-argument relations, discourse functions, etc.

A *treebank*, in contrast, is a corpus annotated at levels beyond the single word. Treebanks

are so named because of the common practice of representing syntactic structure in the form of phrase structure trees, even though syntactic and semantic representations may have other forms. The kind of annotation chosen and the methodology for annotation depend of course on the purpose of the annotation, in other words, what the corpus will be used for. In the rest of this paper, two alternative approaches to syntactic annotation of a learner corpus will be described.

Manual annotation of interlanguage

Dickinson and Ragheb (2009) describe the development of a grammatical annotation scheme for second language learner data. Their aim is to create a resource that will support second language acquisition research, and they formulate their goal as follows: "What needs to be described is *interlanguage*, the in-progress language of learners which is a linguistic system in its own right, without focusing on errors." This is an ambitious and maybe somewhat premature goal since it presupposes systematicity in interlanguage, whereas the level of systematicity in interlanguage is an interesting research goal in its own right (Tenfjord, 1983, p. 10).

Unlike the target language, for which intersubjective norms govern grammaticality, interlanguages are individual (Tenfjord, 1997, p. 8). Therefore, it may be difficult to assign language learner errors clearly to either competence or performance. Other issues related to the status of interlanguage (e.g. Henderson, 1985) since the term was first introduced into SLA research by Selinker (1972) should also be taken into account before embarking on any attempt to annotate interlanguage as a linguistic system.

Dickinson and Ragheb say that they are not aware of any previous attempts to syntactically annotate interlanguage, although they do refer to work on automatic parsing of learner data with the aim of detecting errors, such as Menzel and Schröder (1999), who parse learner language for the purpose of diagnosis in tutoring systems. They are skeptical to the resulting dependency structures in that work, since "it is not clear what exactly the surface syntax is encoding, as the parse is based on a model of native language." They say further that "it is unlikely that surface dependencies (or constituencies) capture the full set of syntactic facts employed by a learner."

Thus Dickinson and Ragheb want to design an annotation scheme that neither involves error annotation nor rests on an L2 description. Although they acknowledge that the study of errors can be interesting, they say that error annotation is not so useful for studying properties such as fluency, complexity and stage of acquisition. They followed several general principles in developing their annotation. The most important of these is probably that learner language should be annotated "as is", without marking errors or positing "empty elements or corrected forms". Dickinson and Ragheb (2009, p. 61) "want to make as few claims as possible about what the intended meaning of the learner is, aiming only at an adequate description of the learners' interlanguage, from which researchers can draw their own conclusions".

Their goal is thus quite ambitious, since their intention is to code the linguistic system underlying the learners' language rather than that of the L2. As we will see, however, their annotation scheme rests on L2 properties and on the interpretation of the learner language in relation to those properties. Dickinson and Ragheb's annotation scheme consists of two main parts, POS annotation at word level, and dependency annotation to represent syntactic relations. These are described in the following sections.

Part of speech annotation

Dickinson and Ragheb lemmatize each word in the corpus, including spelling mistakes. In order to do this they must of course interpret spelling errors and relate them to intended words, so that even at this simple level, interpretation is unavoidable. The POS tagging is then done using the tagset from the SUSANNE corpus (Sampson, 1995). They modify the tagset by splitting the POS annotation into two parts: one tag is based on the linguistic form of the word, while the other refers to its syntactic use. For most words, the two tags will be the same, but when there is some anomaly, the tags will be different. They provide example 1 to illustrate how the tagging is done.

(1) Tin Toy can **makes** different music **sound**.

The verb form *makes* is assigned the tag *VVZt* (third person singular present tense) as well as *VV0t* (baseform verb). The first tag is meant to account for the actual occurring form, while the second provides the syntactic function. Despite Dickinson and Ragheb's stated goal of annotating interlanguage without reference to errors, one way of interpreting this annotation scheme is precisely that it is error encoding. In this example, the third person singular present tense form is marked as being the wrong form in a syntactic position that calls for the base form of the verb.

As concerns the word *sound* in the same example, the authors state that it clearly has the form of a singular noun, but that the learner "may be using this form as either a singular or plural noun". Since *different music sound* is not a well-formed nominal phrase in the L2, they see two possible interpretations. Either *sound* was intended to be a plural, or it was intended as a singular, but the phrase is anomalous since there is no determiner. In order to avoid making a decision on the intended reading, Dickinson and Ragheb provide both the form tag for a singular noun, *NN1c*, and an underspecified use tag, *NN*. However, we find this example also to be ambiguous between a noun and a verb reading. In fact, with the verb reading there is nothing anomalous about the string *different music sound*. Also in this case, the authors have perhaps interpreted more than they intended to.

In some cases they define new tags to account for learner language, as illustrated in example 2. Here they propose a compound tag because *adram* "seems to be a blend of *a drum*."

(2) The tin toy had **adram** and a acordion.

It is not convincing that this should be a special feature of learner language; it could also simply be considered a spelling error, more precisely perhaps one typographical error (the missing space) and one orthographical error (the substitution of one vowel for another).

A more controversial case of a new tag is proposed for the following example.

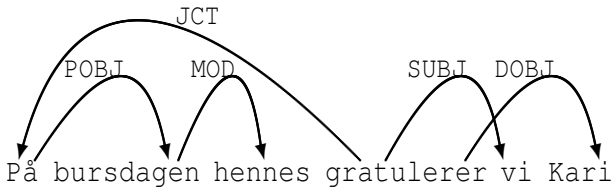
(3) The child **follow** him.

Dickinson and Ragheb say this about example 3: "In this case, we do not know the specific POS use of *follow*; we only know that it is tensed." This sentence cannot be successfully annotated using the two-tag strategy outlined above, since it is not clear what the intended syntactic usage is in the L2. Therefore, a new underspecified tensed verb tag *VVTt* is introduced that identifies *follow* as a tensed verb without saying which tense it expresses. It is not obvious, however, that this verb form, identical to the infinitive, should be considered tensed. Many languages do not grammaticalize time reference as tense, an example being Vietnamese. Tenfjord (1997) has shown that Vietnamese learners of Norwegian acquire tense very late, if at all; they use untensed verb forms to a great degree where the L2 requires tensed forms, sometimes long after they have acquired the perfect, which is grammaticalized in Vietnamese. In this case, then, we believe that Dickinson and Ragheb are attributing an L2 category to the learner language that it doesn't necessarily have.

Dependency annotation

For syntactic annotation, Dickinson and Ragheb mark dependency relations between words. This methodology, stemming from dependency grammar (Tesnière, 1959), marks grammatical relations between individual words in a sentence, rather than grouping words into phrases as in the Chomskyan tradition. This kind of analysis is illustrated in 5, where we show a possible dependency annotation for the constructed example sentence in 4.

(4) *På bursdagen hennes gratulerer vi Kari.*
on birthday.the hers congratulate we Kari
"On her birthday we congratulate Kari."

(5)


Dickinson and Ragheb's motivation for using dependency rather than constituency is their claim that this kind of encoding can be done more quickly. They use the grammatical relations

proposed in Sagae et al. (2007) in connection with CHILDES and encode the dependencies in the format proposed in Buchholz and Marsi (2006).

Despite Dickinson and Ragheb's aim of annotating interlanguage as such, they use clear L2 properties as evidence in determining the usage of a word in a sentence context. In particular, they let standard English word order guide the assignment of grammatical relations, also in the case of mismatches with morphological marking, as in their constructed example 6, or in the case of mismatches involving subcategorization.

(6) **Him** wants to save his life.

The principle put forward by Dickinson and Ragheb (2009, p. 63) to "place a greater emphasis on word order, or positional information for determining grammatical relations", which favors assigning subject function to *him* in example 6, can hardly be said to serve annotation of learner language as such. It is clearly pointing to an interpretation, and in addition, an interpretation that is grounded in target language rather than in interlanguage characteristics.

With respect to example 7, Dickinson and Ragheb (2009, p. 64) say that the "word *dull* (assuming its intended form *doll*) is ambiguous: it could be an object of *escape* (with a missing subject), or it could be the subject in the wrong location."

(7) ...escape the dull [doll]

The authors disprefer the possibility of underspecifying the dependency label and choosing 'arg' rather than subject or object. Rather, they suggest a distinction between 'surface' and 'underlying' dependencies to account for such 'learner ambiguities'. Thus they assign *the dull* a surface dependency as object and an underlying dependency as subject. However, we think this approach does not take into account other possibilities. It is conceivable, for instance, that the learner's interlanguage has a systematic VS word order. In that case, it might not be warranted to assign object function to *the dull* at any level of description, except for the purpose of error coding, as Dickinson and Ragheb (2009, pp. 68) suggest: "by annotating the layers separately, we point to the error".

In conclusion, they want to avoid interpreting the output of the learners, but in the treatment of their examples it is clear that they are interpreting. They also want to avoid error annotation, but they do project the syntax of the correct L2 onto the learner's utterances and annotate anomalies. In addition, Dickinson and Ragheb (2009, p. 69) resort to an explicit error tag *JCT+* for coding the anomalous word order of adjuncts, which is not treated through the dependency scheme.

The dependency annotation is generally considered rather easy to do, since the annotator always takes into account only a head-dependent relation between two words. However, despite the intention of Dickinson and Ragheb (2009, p. 66) "to make as few decisions as possible", the

manual annotation which they propose requires annotator decisions for every word and relation between words.

Automatic parsing with an L2 grammar

An alternative approach to syntactic annotation of learner language consists of automatic parsing with a grammar of the L2. The ultimate objectives of this approach are to a large extent similar to Dickinson and Ragheb's, except that it explicitly admits to interpretation and error coding, and has the advantage of having a grammar of the L2 behind it, which promotes a highly consistent annotation.

Automatic parsing of original learner language with an L2 grammar would, however, cause the parser to fail for utterances that violate L2 grammar rules. Parser failures could in some cases indicate where a learner error occurs, but this would not necessarily provide a good characterization of the anomaly. A more realistic approach to automatic syntactic annotation of a learner corpus would be analysis of the 'correct' version, as the term is used in ASK, i.e. the version which has been manually constructed by interpreting the L2 texts and rendering them in correct target language.

The goal of such an annotation effort would be more modest: we do not propose annotating 'interlanguage' as such, but rather to provide a treebank which facilitates research by making it possible to search for the syntactic contexts in which learners' errors occur. In this respect, we consider interpretation of the intended reading not only permissible, but even necessary for a treebank. If interpretation were totally excluded, it would not be possible to resolve any ambiguities, thereby rendering the syntactic annotation nearly useless for effective search in the corpus.

We propose the use of NorGram (Dyvik, 2000), a computational grammar for Norwegian based on Lexical-Functional Grammar (LFG) (Bresnan, 2001). With this grammar, a corpus can be annotated with detailed information at three levels of structure:

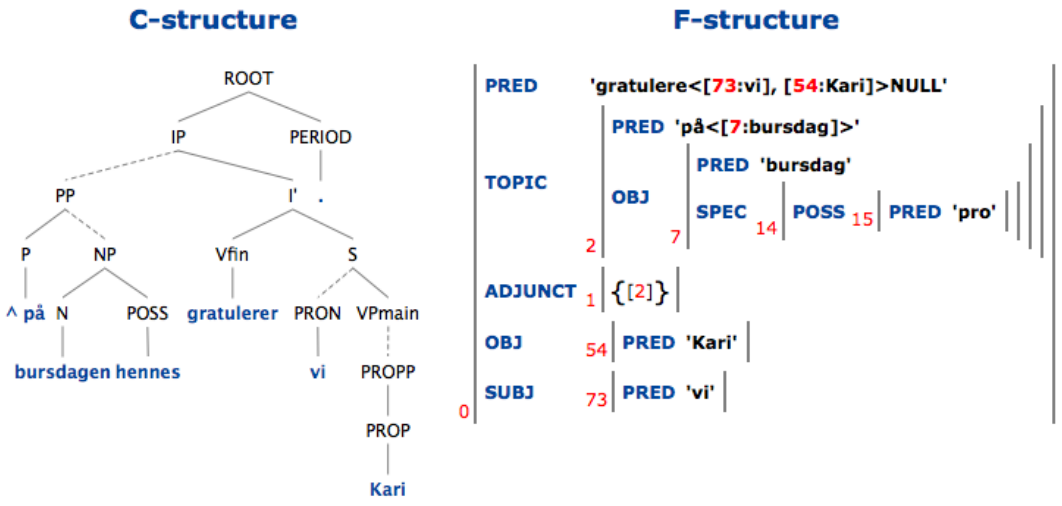
1. constituent structure (c-structure), capturing hierarchical groupings in terms of phrase structure;

2. functional structure (f-structure), representing predicate-argument relations, syntactic relations and syntactic features in terms of recursive attribute-value pairs;

3. MRS-structure, a semantic structure based on Minimal Recursion Semantics (Copestake et al., 2005).

These levels of structure provide complementary and consistent information. An example of a c-structure and an f-structure, with some detail omitted, is provided in 8 for the same sentence as in 4.

(8)



These structures make it easy to retrieve whole phrases and their properties. A recent research project, for instance, demonstrated a clear need to be able to search for all complex noun phrases in a learner corpus which require gender agreement (Ragnhildstveit, 2009, p. 77). Another project (Nordanger, 2009) faced a similar problem when attempting to identify NPs without adjuncts. This information is very difficult to retrieve in a corpus tagged at word level, but is easy to find in a treebank.

The LFG PARSEBANKER (Rosén et al., 2009) is a tool which is well suited for building a treebank as a parsed corpus (often called a *parsebank*). The first step consists of the automatic parsing of a corpus and the storage of all analyses in a database. With the help of an efficient disambiguation procedure, a human annotator chooses the best analysis, which is then recorded (Rosén et al., 2007).

A treebank based on a 'correct' version of a learner corpus may be useful in various respects, even if learner language as such is not annotated directly. On the one hand, the 'correct' treebank can provide the syntactic context in which errors occur. This allows for more specific searches, in particular for a combination of a particular error code in the corpus and a syntactic pattern in the 'correct' version. On the other hand, the 'correct' treebank still contains all the learners' correct language use as well. A good characterization of the complexity of L2 at specific learning stages can be obtained from the annotation in the 'correct' treebank together with the error annotation. In that respect, it will also be possible to search for constructions where certain error codes do not occur, which may indicate correctly acquired grammar rules.

It is interesting, for instance, to investigate hypotheses about possible factors determining whether inversion is correctly applied or is missing (Johansen, 2008). Some hypotheses might involve the syntactic function of the preposed constituent, the category of the subject (e.g. pronoun or full-fledged NP), the frequency of the finite verb, etc. (Hagen, 1992). The error code in ASK for missing inversion does not in itself give an indication of the context in which this

error occurs, nor does it provide an indication of where inversion is correctly applied. Given a treebank of the 'correct' version, however, a thorough and efficient investigation of such hypotheses may be supported by selecting sentences with inversion in the 'correct' version and investigating the co-occurrence of the presence or absence of the error code with relevant syntactic characteristics.

Such investigations are quite feasible with the extensive search possibilities in the LFG PARSEBANKER, which allows for combinations of structural characteristics at different levels (Rosén et al., 2009). It would be straightforward to search, for instance, for Norwegian sentences with correct or missing inversion, starting with different kinds of modifiers, having NPs of varying complexity as subjects. Other features such as mood (declarative, imperative or interrogative), voice (active or passive), etc. can also be included among the search criteria.

Using the 'correct' version, based on probable intended meanings, means that there will be a better chance of obtaining good analyses than might be possible with the original text. One of the purposes of the 'correct' version in ASK has in fact been to allow parsing. However, even after normalization to correct L2 syntax, there will certainly still be challenges to automatic parsing. Various kinds of errors may remain undetected in the original phase of error annotation. The LFG PARSEBANKER offers a possibility for the annotator to capture and correct word level errors, such as typographical errors, so that the parser can still find the correct analysis (Rosén, 2008). At the same time, the system does not remove the original text, but retains it. Also, in possible cases of missed errors in the corpus, the annotator has the possibility of marking parts in the utterance to be ignored by the parser. In addition, the LFG PARSEBANKER provides the annotator with the option of manually segmenting utterances with informal constructions that would be problematic for the parser, for instance in the case of run-on sentences. If the parser encounters passages not covered by the grammar, it will still come up with a *fragment analysis*, i.e. will analyze all fragments to which it can assign a partial analysis (Rosén and De Smedt, 2007). These options, which are already implemented, could enhance the success rate of automatic parsing for the purpose we have in mind.

Conclusion

While we agree with Dickinson and Ragheb (2009) in that a syntactic characterization of authentic learners' language would be very useful for research purposes, we do not think that such a characterization can be achieved in a straightforward manner without considerable interpretation of the utterances that learners produce. Nor is it realistic to exclude any reference to normal L2 during the annotation.

We have shown that while Dickinson and Ragheb (2009, p. 61) "try to annotate language *as is*, i.e. annotate only what is there", they have difficulties in adhering to this principle. Every

annotation, whether lemmatization, POS tagging, or the assignment of syntactic relations, involves a process of interpretation. We have also shown that their treatment of examples is not completely in line with their attempt "to give the learner the benefit of the doubt" (ibid.) since in several cases they overlook possible readings.

Dickinson and Ragheb deserve credit for proposing an annotation scheme for interlanguage. Their scheme is a step towards being able to search for word tags and dependencies in interlanguage corpora. Dickinson and Ragheb (2009, p. 68) also discuss how "mismatches between annotation levels point to errors." In that respect, their scheme is, however, circular to the extent that the mismatches are both the result of the identification of anomalies and the basis for finding errors. We agree with Tenfjord et al. (2006a) that it is unrealistic to describe interlanguage without reference to the L2 and that error annotation does have methodological value.

As an alternative, we have described a more transparent method based on a mature treebanking technology involving automatic parsing and computer-aided manual disambiguation. This approach, as implemented in the LFG PARSEBANKER, has the advantage of achieving a high degree of consistency and grammatical detail. If a normalized version of a learner corpus is available, as is the case in ASK, then annotation of this 'correct' version is proposed as efficient and realistic because a normal grammar of the L2 can be reused.

We do not intend a treebank of a 'correct' version by itself to be a substitute for the annotation of original learner utterances. We only intend it to be a useful tool that complements the annotation of errors. Together, these could provide a rich, detailed characterization of learner language. In particular, a treebank created in the way we have sketched would be able to exploit the extensive search possibilities in the LFG PARSEBANKER in order to find specific construction types, their frequencies, and possible correlations. This would allow for targeted searches for specific error codes in the corpus material, at the same time taking into account their full syntactic context. It would also allow searches for constructions without errors and could be a valuable tool for exploring many aspects of learner language, for example avoidance strategies.

References

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Malden, MA: Blackwell.

Buchholz, Sabine and Marsi, Erwin. 2006. CoNLL-X shared task on Multilingual Dependency Parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 149–164, Association for Computational Linguistics, New York City.

Copestake, Ann, Flickinger, Dan, Pollard, Carl and Sag, Ivan A. 2005. Minimal Recursion Semantics: An Introduction. *Journal of Research on Language and Computation* 3(4), 281–332.

Dickinson, Markus and Ragheb, Marwa. 2009. Dependency annotation for learner corpora.

In Marco Passarotti, Adam Przepiórkowski, Savina Raynaud and Frank Van Eynde (eds.), *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*, pages 59–70, Milan, Italy: EDUCatt.

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 Øivin Andersen, Kjersti Fløttum and Torodd Kinn (eds.), *Menneske, språk og felleskap*, Novus forlag.

Granger, Sylviane. 2008. Learner corpora. In Anke Lüdeling and Merja Kytö (eds.), *Corpus linguistics: an international handbook*, volume 1, pages 259–275, Berlin and New York: Walter de Gruyter.

Hagen, Jon Erik. 1992. Feilinvertering, overinvertering og underinvertering [Incorrect inversion, overinversion and underinversion]. *NOA* 15, 27–38.

Henderson, Michael M. T. 1985. The interlanguage notion. *Journal of Modern Language Learning* 21, 23–27.

Johansen, Hilde. 2008. Inversjon i norsk innlærerspråk - En undersøkelse av variasjonsmønstre i skrevne tekster [Inversion in norwegian learners' language - An investigation of variation patterns in written texts]. *NOA* 24(2), 50–71.

Menzel, Wolfgang and Schröder, Ingo. 1999. Error Diagnosis for Language Learning Systems. In *ReCALL: the Journal of EUROCALL*, volume 11, pages 20–30.

Nordanger, Marte. 2009. *Keiserens nye klær? Lingvistisk og konseptuell transfer i markeringen av grammatikalisert definitt referanse i russiskspråklige og engelskspråkliges norske mellomspråk - en studie basert på ASK [The emperor's new clothes - Linguistic and conceptual transfer in the marking of grammaticalized definite reference in the Norwegian interlanguage of speakers of Russian and English - A study based on ASK]*. Masters Thesis, University of Bergen, Section for Scandinavian Language and Literature.

Ragnhildstveit, Silje. 2009. *Genustildeling og morsmålstransfer i norsk mellomspråk: En korpusbasert studie [Gender assignment and first language transfer i Norwegian interlanguage: A corpus-based study]*. Masters Thesis, University of Bergen.

Rosén, Victoria. 2008. Mot en trebank for talespråk. In Janne Bondi Johannessen and Kristin Hagen (eds.), *Språk i Oslo. Ny forskning omkring talespråk*, pages 214–225, Oslo: Novus forlag.

Rosén, Victoria and De Smedt, Koenraad. 2007. Theoretically Motivated Treebank Coverage. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA-2007)*, pages 152–159, Tartu: Tartu University Library.

Rosén, Victoria, Meurer, Paul and De Smedt, Koenraad. 2007. Designing and Implementing Discriminants for LFG Grammars. In Tracy Holloway King and Miriam Butt (eds.), *The*

Proceedings of the LFG '07 Conference, pages 397–417, Stanford: CSLI Publications.

Rosén, Victoria, Meurer, Paul and De Smedt, Koenraad. 2009. LFG Parsebanker: A Toolkit for Building and Searching a Treebank as a Parsed Corpus. In Frank Van Eynde, Anette Frank, Gertjan van Noord and Koenraad De Smedt (eds.), *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht: LOT.

Sagae, Kenji, Davis, Eric, Lavie, Alon, MacWhinney, Brian and Wintner, Shuly. 2007. High Accuracy Annotation and Parsing of CHILDES Transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32, Prague.

Sampson, Geoffrey. 1995. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press.

Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics* 10, 209–241.

Tenfjord, Kari. 1983. *Systematisk, variert, tilfeldig: Ein kontrastivt basert analyse av mellomspråket til ei gruppe vietnamesiske ungdommar [Systematic, varied, arbitrary: A contrastively based analysis of the interlanguage of a group of Vietnamese youths]*. Masters Thesis, University of Bergen.

Tenfjord, Kari. 1997. *Å ha en fortid på vietnamesisk: En kasusstudie av fire vietnamesiske språkinnlæreres utvikling av grammatisk fortidsreferanse og perfektum [Having a past in Vietnamese: A case study of four Vietnamese langauge learners' development of grammatical past reference and perfect]*. Ph. D.thesis, University of Bergen.

Tenfjord, Kari, Hagen, Jon Erik and Johansen, Hilde. 2006a. The hows and whys of coding categories in a learner corpus (or "How and why an error-tagged Learner corpus is not ipso facto one big comparative fallacy"). *Rivista di Psicolinguistica Applicata* 6(3), 93–108.

Tenfjord, Kari, Meurer, Paul and Hofland, Knut. 2006b. The ASK corpus: A language learner corpus of Norwegian as a second language. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1821–1824.

Tesnière, Lucien. 1959. *Eléments de syntaxe structurale*. Paris: Editions Klincksieck.