The Interplay Between Lexical and Syntactic Resources in Incremental Parsebanking

Victoria Rosén[1], **Petter Haugereid**[2], **Martha Thunes**[3], **Gyri Smørdal Losnegaard**[4], **Helge Dyvik**[5], **Paul Meurer**[6]

University of Bergen[1,2,3,4,5], Uni Research Computing[1,5,6]
Bergen, Norway
victoria@uib.no[1], petter.haugereid@lle.uib.no[2], martha.thunes@lle.uib.no[3], gyri.losnegaard@lle.uib.no[4]
helge.dyvik@lle.uib.no[5], paul.meurer@uni.no[6]

Abstract

Automatic syntactic analysis of a corpus requires detailed lexical and morphological information that cannot always be harvested from traditional dictionaries. In building the INESS Norwegian treebank, it is often the case that necessary lexical information is missing in the morphology or lexicon. The approach used to build the treebank is incremental parsebanking; a corpus is parsed with an existing grammar, and the analyses are efficiently disambiguated by annotators. When the intended analysis is unavailable after parsing, the reason is often that necessary information is not available in the lexicon. INESS has therefore implemented a text preprocessing interface where annotators can enter unrecognized words before parsing. This may concern words that are unknown to the morphology and/or lexicon, and also words that are known, but for which important information is missing. When this information is added, either during text preprocessing or during disambiguation, the result is that after reparsing the intended analysis can be chosen and stored in the treebank. The lexical information added to the lexicon in this way may be of great interest both to lexicographers and to other language technology efforts, and the enriched lexical resource being developed will be made available at the end of the project.

Keywords: treebanking, INESS, lexicon

1. Introduction

Incremental parsebanking presents a unique opportunity for enrichment of the lexicon. It provides a useful context for supplementing the information provided in lexical resources derived from traditional dictionaries, thus helping to overcome their limitations.

The INESS project (Infrastructure for the Exploration of Syntax and Semantics) is developing a large parsebank for Norwegian.[1] In the process, an existing grammar and lexicon for Norwegian are further developed in tandem. Since the grammar requires quite detailed morphosyntactic information in order to provide an analysis, the lexicon must be syntactically well informed; feedback from the parsebanking process results in a considerable enrichment of the original lexical resource.

In the following, we will first discuss how the syntax and lexicon mutually inform each other in our approach. In section 3. the interface for preprocessing texts will be described. The treatment of unknown words will be illustrated in section 4. In section 5. the incremental parsebanking approach in INESS is briefly described. Finally, in section 6., we will present various kinds of missing or incorrect information in lexical resources and show how this may be remedied.

2. The interplay between syntax and lexicon

NorGram is a hand-written computational grammar for Norwegian (Dyvik, 2000; Butt et al., 2002). It is written in the Lexical Functional Grammar (LFG) framework (Bresnan, 2001; Dalrymple, 2001). The Xerox Linguistics Environment (XLE) is used for grammar development and parsing (Maxwell and Kaplan, 1993). NorGram has been used in several language technology projects, and its main lexicon has been the NorKompLeks electronic lexicon (Nordgård, 2000). This lexicon is an adapted version of *Bokmålsordboka*, a dictionary of Norwegian Bokmål (Landrø and Wangensteen, 1993), and *Nynorskordboka*, a dictionary of Norwegian Nynorsk (Hovdenak et al., 1986).

NorGram provides deep syntactic analysis on two levels: constituent structure (c-structure) and functional structure (f-structure). The c-structure is a phrase structure tree showing the linear and hierarchical organization of the phrasal constituents in the sentence. The f-structure is an attribute–value matrix showing grammatical functions and features.

In LFG, the syntax and the lexicon have an important interaction with each other especially in the treatment of predicate–argument structure. The lexical entry for each verb must specify which arguments a verb requires. If the sentence lacks syntactic arguments which the verb specifies, or if the sentence contains syntactic arguments which the verb does not specify, no grammatical analysis will be produced. For example, in a transitive sentence, the lexical entry for the verb must specify that the verb can take an object.

3. Text preprocessing

An important source of texts for the INESS Norwegian treebank is a large repository of OCR-read fiction texts supplied by the National Library of Norway. Because OCR software makes certain errors, such as misinterpreting characters, omitting text, or inserting unwanted material, the documents must be preprocessed before syntactic parsing. Moreover, when a corpus is parsed, there will always be words that are unknown to the morphology and/or the lexicon. INESS has developed an intelligent preprocessing in-

[1] http://clarino.uib.no/iness

terface which facilitates efficient text cleanup and the treatment of unknown word forms (Rosén et al., 2012b).

Text cleanup involves for example removing superfluous material that does not belong to the text, joining parts of sentences that have erroneously been split, and adding punctuation where it is missing. The interface offers practical editing functions for these operations.

After text cleanup, the annotators process word forms that have not been automatically recognized. The preprocessing interface presents a list of unknown words. Some of these are the result of OCR errors, and some are simply typos. Other frequent types of unrecognized word forms are productive compounds, multiword expressions, named entities, foreign words, neologisms, interjections, dialect words, and systematic, or intended, misspellings. It is important that the annotator observes the difference between typos, or unintentional misspellings, which must be corrected along with OCR errors, and nonstandard word forms, which are not to be changed.

We distinguish between three main classes of nonstandard word forms. These are systematic misspellings, archaic word forms, and nonstandard forms that can be ascribed to a particular dialect, technolect, sociolect, or other language variety. Systematic misspellings are, typically, not just incidental typos, but forms produced regularly by an author. During preprocessing unrecognized forms of these types are left unchanged because correcting them would be to interfere with actual language use.

The important common denominator of all types of unrecognized words which are not to be corrected is that while these forms fall outside standard dictionaries, it is a prerequisite for successful parsing that they are included in our lexicon. Since one unknown word may result in the parser not returning an analysis, it is important to recognize and properly treat as many such words as possible.

4. The recognition of unknown words during preprocessing

The preprocessing interface allows the annotators to add unrecognized words to the lexicon so that they will not cause parsing failures. Noninflecting words like named entities and interjections are entered as simple paradigms, with a given category assigned to each entry. Inflecting words belonging to the open lexical classes are entered as complex paradigms, and the annotator must specify an inflectional pattern for each new entry. Verbs must also be assigned subcategorization frames necessary for parsing. When a word is unrecognized because of nonstandard spelling, the annotator must consider whether the spelling deviation concerns the stem or an inflection. Variant stems are entered as paradigms associated with an existing standard paradigm, and variant inflectional forms are registered as deviations of individual, standard inflectional forms. In order to add unrecognized words to the lexicon in an efficient way, the annotator makes use of a set of predefined options in the preprocessing interface. Each option corresponds to a certain type of entry. Most of these types can be entered by a single mouse click, while the recording of paradigms and variant inflectional forms requires a few more steps. Table 1 presents an overview of the number of unrecognized words that have been extracted through preprocessing of a corpus of about 29 million words. Among these words, members of the open lexical classes (nouns, verbs, adjectives, and adverbs) are the most frequent type of words added through preprocessing (39,2% of all entries). These are given as the category *paradigm* in table 1, and within the class of paradigms, more than half of them were compounds (6,595 entries). Table 2 lists some of the most frequent compound types added through preprocessing in this study. Prior to preprocessing, an automatic compound analyzer is run on the text in order to identify compounds that are not already in the lexicon. The analyzer checks for a certain set of patterns, and compounds that are not recognized are presented to the annotator as unrecognized words.

Category	Instances
Paradigm (open word class)	12477
Last name	3335
Place name	3306
Organization or brand name	2479
Unclassified	1847
Foreign expression	1505
Miscellaneous name	1463
Variant inflectional form	1310
Person name	1306
Title	1053
Interjection	849
First name, masculine	839
First name, feminine	590
Taxon name	72
Total	31828

Table 1: Overview of the various types of unrecognized words added through preprocessing.

Compound type	Example	Instances
noun+noun	*appelsin+te* ‘orange-tea’	3110
noun+adj	*avis+grå* ‘newspaper-gray’	1284
adj+adj	*blå+brun* ‘blue-brown’	444
adj+noun	*fin+kåpe* ‘nice-coat’	230
prep+noun	*av+knapp* ‘off-button’	160
adj+verb	*blek+pudre* ‘pale-powder’	140
prep+verb	*av+beite* ‘off-graze’	137
verb+noun	*ete+fest* ‘eat-party’	116
Others		1196
Total		6595

Table 2: Overview of some of the most common compound types added through preprocessing.

The screenshot in figure 1 illustrates how the unknown compound *gulblank* ‘yellow-shiny’ is added to the lexicon by the annotator. As the base form is entered, the annotator marks the internal structure of the compound by separating the first and second element by the character +. Moreover, if the lexical class of the first element is a category other than noun, this category is entered in parentheses (in this case *adjective*). When the base form has been typed in, the annotator must specify an inflectional paradigm for the new lemma, either by typing in the base form of an existing lemma with matching inflection (in this case the adjective *blank*), or by selecting one from a set of potentially matching lemmas proposed by the interface. An inflectional paradigm must be specified for all paradigm entries, whether they are compounds or not.

The motivation for analyzing unrecognized compounds in this way (by registering the part of speech also of the first part) is to be able to discover frequent compound elements and compound types that are not already accounted for by the compound analyzer. The noun+noun type is very frequent, and is normally handled by the analyzer. The example of this type in table 2, *appelsin-te*, was not recognized because the second element is a two-letter word. Allowing compound consituents of three letters or less is generally considered a risk in automatic compound analysis; if such short constituents are allowed in general, practically any typo or misspelled word could be erroneously analyzed as a compound.

Of the types included in table 2, the following are currently handled by the compound analyzer: noun+noun, noun+adj, adj+noun, adj+verb and verb+noun. Some of these combinations have certain constraints imposed on them. For noun+adj compounds, only a few nouns that occur frequently as the first element in compounds are allowed; examples are *døds* 'death', *kjempe* 'giant', *drit* 'shit', and *rekord* 'record'. For adj+verb compounds, the verb may only be a past participle. These constraints explain why the compounds *avisgrå* 'newspaper-gray' and *blekpudre* 'pale-powder' were not recognized. The types adj+adj, prep+noun, and prep+verb are currently not allowed at all. Studying the individual examples in the different categories will help to determine if new types should be added to the compound analyzer, or whether some particularly frequent elements should be allowed.

Table 1 also shows that the second most frequent type of unrecognized words in this study is named entities. Among these, last names, place names, and organization or brand names are very common.

The next category listed in table 1 is *unclassified*. This is a residual class used for unrecognized words that fall outside the set of predefined options available in the preprocessing interface, typically because they have some kind of morphological or morphosyntactic property which does not match any of the available categories. This is the case for certain word forms involving clitics, like *n'Oscar*, which is a contraction of the pronoun *han* 'he' and the name *Oscar*. Such forms are entered as unclassified, pending further treatment by the grammar developer. Another type of word that must be entered as unclassified is compounds which can be regarded as products of syntactic processes, such as *gamlebilen* 'the old car'. The first element, *gamle*, is an adjective inflected in the singular definite form, and the second element, *bilen*, is a noun, also with a singular definite inflection. This compound does not have a normal inflectional paradigm; it will not occur in the plural, or in the indefinite, because it is a contracted form of a syntactic phrase *den gamle bilen* 'the old car'.

Foreign words are often used in Norwegian sentences. Sometimes they are spontaneous uses of a word from another language, most often English. Other times they are well-established in Norwegian, but have not yet made their way into standard dictionaries. An example of a spontaneously used English word is shown in (1).

(1) *«Jeg skulle ikke være noen alien for deg,»*
I should not be some alien for you
sa Auguste.
said Auguste
' "I'm not really an alien for you," said Auguste.'

Example (2) contains both the English loan *air conditioning* and the named entity *American Bar*.

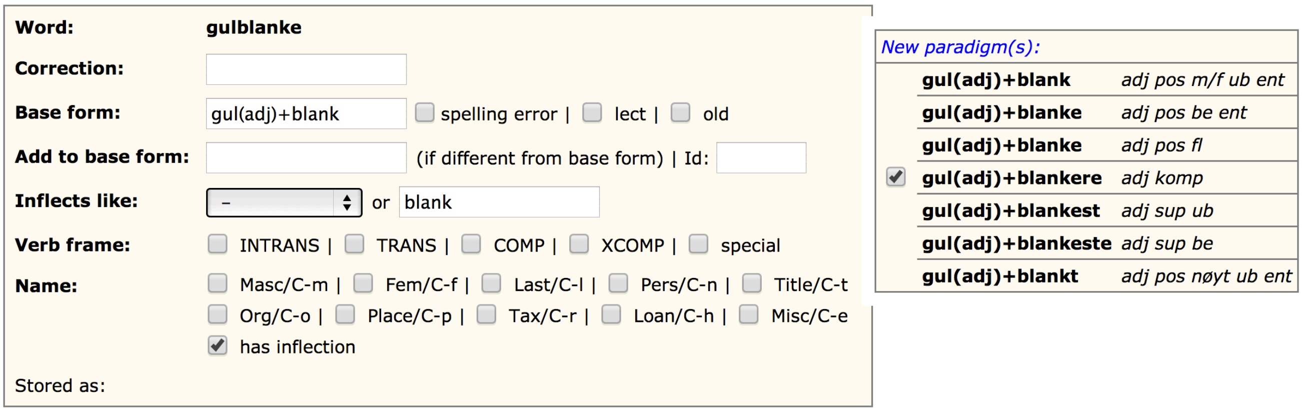
(2) *Han gikk inn på American Bar, som*
he went in on American Bar which
reklamerte med air conditioning.
advertised with air conditioning
'He entered the American Bar, which boasted air conditioning.'

Missing lexical entries like this are easily added to the lexicon when they are identified in the preprocessing step. In this case, *American Bar* was entered as a named entity of the category organization name, and *alien* and *air conditioning* were entered as loans.

A particularly productive part of speech is interjections; especially writers of fiction are very creative in the way in which they write interjections. *Bokmålsordboka* has an entry for the interjection *hysj* 'hush' which includes also the alternative spelling *hyss*. There are several occurrences of this interjection in the fiction texts of the INESS treebank, and many of them do not have either of the two standard spellings. The following eight variants of *hysj/hyss* have been registered until now: *hysjjj*, *hyssj*, *hysssj*, *hyssssjjj*, *hyssssssj*, *hyssssssjjj*, *hysst*, *hyyyysssjjj*. This shows that the spelling of this interjection is unpredictable and to a large extent determined by the way in which an author chooses to express it in a given context. For parsebanking purposes, the challenge is that each time a new spelling is encountered, it is displayed in the preprocessing interface as an unknown word. The INESS interface makes it possible for annotators to add new variant spellings to a single lemma in the lexicon. In this way each extracted variant can eventually be recognized during parsing.

It can often be justified to add misspellings to the lexicon and/or morphological analyzer. An author can for instance use a creative spelling to imitate a certain dialect or pronunciation. An example from the INESS parsebank is *mordern* 'the murderer', instead of the standard form *morderen*. The elided vowel is imitative of a certain accent. The annotator enters the form as a new inflectional variant by indicating in the preprocessing interface that it shares the same lemma and the same inflectional features as *morderen*.

Thus, as the annotator processes the unrecognized words in a document, new lexicon information is compiled, and before the text is syntactically parsed, this new information is added to the lexical resources exploited by the parser.



Word:	**gulblanke**
Correction:	
Base form:	gul(adj)+blank □ spelling error | □ lect | □ old
Add to base form:	(if different from base form) | Id:
Inflects like:	– or blank
Verb frame:	□ INTRANS | □ TRANS | □ COMP | □ XCOMP | □ special
Name:	□ Masc/C-m | □ Fem/C-f | □ Last/C-l | □ Pers/C-n | □ Title/C-t □ Org/C-o | □ Place/C-p | □ Tax/C-r | □ Loan/C-h | □ Misc/C-e ☑ has inflection
Stored as:	

New paradigm(s):

	gul(adj)+blank	*adj pos m/f ub ent*
	gul(adj)+blanke	*adj pos be ent*
	gul(adj)+blanke	*adj pos fl*
☑	**gul(adj)+blankere**	*adj komp*
	gul(adj)+blankest	*adj sup ub*
	gul(adj)+blankeste	*adj sup be*
	gul(adj)+blankt	*adj pos nøyt ub ent*

Figure 1: Interface for adding unknown words during preprocessing

Type of lexicon update	No.
Verbs:	
new MWE frame	237
new intransitive reading	46
new inquit reading	30
new transitive reading (incl. ditrans.)	22
new intransitive with expletive subj.	15
miscellaneous new verb frames	12
Adverbs and prepositions:	
new adverb readings	47
new preposition	3
Nouns:	
new mass reading	47
new MWE frame	27
added count noun	17
new title reading	6
Adjectives:	
new MWE frame	12
new adjective	2

Table 3: Overview of lexicon updates made by annotators.

5. Incremental parsebanking in INESS

In our parsebanking approach, the output of the parser is semi-automatically disambiguated by annotators. Thus annotators never manually edit an analysis, but they verify if the analysis produced by the parser is correct, or they choose the correct analysis if several possible analyses are produced. The parsebanking system automatically detects discriminants which help the annotators to efficiently distinguish between possibly many proposed analyses (Rosén et al., 2012a; Rosén et al., 2009; Rosén et al., 2007).

The advantage of our parsebanking approach is that the grammar will always be fully compatible with the treebank. Thus, treebanks constructed in this way achieve a very high level of consistency. Also, the approach is scalable by using stochastic disambiguation to parse new texts fully automatically. However, only sentences that are grammatical—according to the current grammar—will be fully analyzed, while others may receive a fragment parse or may fail to parse.

To the extent that coverage of the grammar needs to be improved, the approach is therefore an incremental one. Annotators signal shortcomings which are followed up by extensions or other changes in the grammar and lexicon, after which the treebank can be reparsed (with cached discriminants to speed up the process).

In INESS we have carried out a detailed study of a small subcorpus in order to find out what the main causes of failed analyses are. We found that 29% of the failed analyses were caused by syntactic problems, while 71% were caused by lexical problems. Of the lexical problems, 41% were caused by missing multiword expressions, whereas 31% were caused by incorrect lexical categories (Losnegaard et al., 2012). This shows that correct lexical information is essential for successful syntactic analysis.

6. Known words with missing or incorrect information

Even though the NorKompLeks lexicon is a rich resource, in parsing we still often find that it lacks lexical information that we need in order to analyze even quite common words. We need, inter alia, lexical category, inflection, subcategorization, countability, compound structure, and multiword expressions. Table 3 gives an overview of the types of lexicon updates made by three annotators while doing parsebanking over a period of about five months.

The NorKompLeks lexicon added subcategorization frames for the verbs in *Bokmålsordboka*. There are, however, many quite common frames that are not included. As shown by table 3, the most frequent type of lexicon update in this study concerns subcategorization frames for verbs, and new verb frames involving multiword expressions (MWEs) account for almost two thirds of these cases.

The effect of updating a verb entry with a new subcategorization frame can be illustrated by example (3) from the INESS treebank. The sentence involves the particle verb *flate ut* 'flatten out'.

(3)	*Fjellet*	*flater*	*ut.*
	mountain.the	flatten	out
	'The mountain flattens out.'		

The Norwegian word form *flate* is categorially ambiguous:

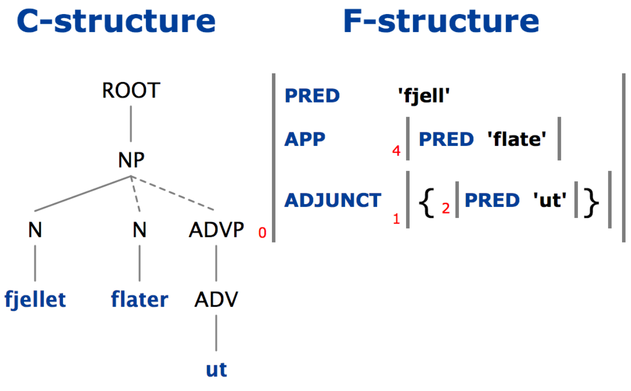


Figure 2: Analysis offered before lexical update

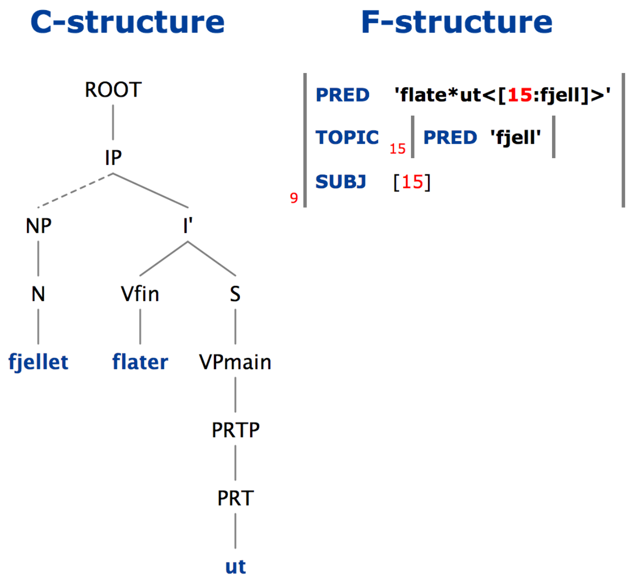


Figure 3: Analysis offered after lexical update

it can be either a verb or a noun. Initially, the lexicon entry for the verb *flate* 'flatten' contained no subcategorization frame covering the MWE *flate ut*; the only verbal frame available required a reflexive element, which does not occur in this sentence. Therefore, the only analysis found by the parser for (3) was that of a noun phrase, where the word form *flater* was analyzed as the plural indefinite of the noun *flate* 'surface' functioning as an apposition to the noun *fjellet* 'the mountain'. Figure 2 shows the c- and f-structures for this analysis of (3). After the annotator added the missing subcategorization frame to the lexicon, the sentence was reparsed. As the c-structure in figure 3 shows, *flater* is now analyzed as a present tense verb (with the lexical category Vfin), and *ut* as a particle (PRT).

Adding this argument frame involves making an addition to a lexical entry, in which an existing template specifying the features of an intransitive verb with a selected particle is called. Figure 4 shows the lexical entry of *flate* with this addition in the second line. The notation {...|...} is a disjunction specifying alternative readings.

```
flate     V XLE { @(V-SUBJ-OBJrefl-PRT flate ut)
                | @(V-SUBJ-PRT flate ut) }; ETC.
```

Figure 4: Lexical entry after the addition of the intransitive frame with particle

```
V-SUBJ-PRT (P PRT) =
       @(CONCAT P `* PRT %FN)
       { (^ PRED)='%FN<(^ SUBJ)>'
         ~(^ PASSIVE)=+
       | { (^ PRED)='%FN<NULL>(^ SUBJ)'
         | (^ PRED)='%FN<(^ OBL-AG)>(^ SUBJ)' }
         (^ PASSIVE)=c +
         (^ PRESENTATIVE-TYPE)=passive
         (^ SUBJ PRON-TYPE)=c expl_
       | (^ PRED)='%FN<(^ OBJ)>(^ SUBJ)'
         (^ PRESENTATIVE)=+
         ~(^ PASSIVE)=+
         (^ SUBJ PRON-TYPE)=c expl_
         ~(^ OBJ DEF)=+ }
       (^ CHECK _PRT-VERB)=+
       (^ PRT-FORM)=c PRT.
```

Figure 5: The template for intransitive particle verbs

The template V-SUBJ-PRT is defined as in figure 5. The first line builds the predicate name 'flate*ut' by concatenation. The following disjunction in the template specifies three alternatives: regular active (*Fjellet flater ut*), impersonal passive (*Det flates ut* 'There is flattening out'), and active presentative (*Det flater ut en fugleflokk* 'There is a flock of birds flattening out').

As table 3 shows, several other types of lexicon updates for verbs are also relatively frequent in this study. We found that new intransitive verb readings were needed in 46 cases, whereas 18 verb frame updates involved adding transitive readings. This is interesting because it may indicate that with respect to verb subcategorization frames, the information available from standard dictionaries does not capture the extent to which verbs with variable argument frames are used intransitively. The sentence in (4) was initially returned by the parser with no analysis, and parsing had failed because the lexicon contained no intransitive reading for *avslå* 'decline'. After this reading was added, the sentence was successfully reparsed.

(4) *Men bestefar avslo.*
but grandfather declined
'But grandfather declined.'

Adding an inquit reading is another frequent type of update in lexical entries for verbs (30 instances). Inquit verbs are verbs of saying and related verbs that may occur in this function, and in the analyzed texts a large variety of verbs are used in inquit clauses. This is not surprising, since the text material is fiction, containing numerous passages of dialogue, as well as inner monologue. The addition of an inquit reading in the lexical entry for a verb involves adding a subcategorization frame specifying that the verb takes a sentence complement as one of its arguments as well as a feature allowing it to occur in the syntactic position typical of inquit verbs.

(5) *Hva mener du med det? stotret hun.*
what mean you with that stammered she

'What do you mean by that? she stammered.'

The sentence in example (5) was initially given a partial analysis by the parser. That is, the word sequences *Hva mener du med det* and *stotret hun* were respectively identified as sentence units, but no complete analysis was found, because the lexicon entry for the verb *stotre* 'stammer' contained only an intransitive reading. An inquit reading was added to the entry, and after reparse the sentence *Hva mener du med det?* was successfully analyzed as a sentential complement to the inquit verb.

Table 3 also shows that lexicon updates involving new readings of adverbs constitute another frequent type in this study (47 occurrences). This illustrates that the lexical category of a given word must often be more fine-grained than what is provided by the lexicon. In the case of adverbs, there is only one large class with the part of speech ADV in the original lexicon. However, different types of adverbs differ considerably in their syntactic distribution, and it is therefore necessary to classify them into subcategories in order to account for this distribution. Our parsing lexicon distinguishes between 22 categories of adverbs based on syntactic position, usually named according to their typical semantic contribution. Thus, between the finite verb and the object there are positions for ADVatt ('attitude adverbs' like *dessverre* 'unfortunately'), ADVprt ('particle adverbs' like *vel* 'I suppose', ADVcmt ('commitment adverbs' like *egentlig* 'actually'), ADVneg ('negation adverbs' like *ikke* 'not'), and others, where there are ordering constraints. Thus, particle adverbs occur before commitment adverbs, which occur before negation adverbs, cf. example (6).

(6) *Jeg har vel egentlig ikke noe*
I have I-suppose actually not something
å legge til.
to lay to
'I actually have nothing to add, I suppose.'

Different classes of degree adverbs are also distinguished, for example ADVdeg ('degree adverbs' like *ganske* 'quite', which modify adjectives) and ADVdegloc ('locational degree adverbs' like *langt* 'far', modifying locative adjuncts), cf. example (7).

(7) *ganske langt fra vannet*
quite far from lake.the
'quite far from the lake'

The category ADV is used for the large class of adverbs which only occur in the VP domain, mostly manner adverbs. When annotators find that the analysis provided by the parser is inadequate, the situation can often be remedied by changing the part of speech from the default category ADV used in NorKompLeks to one of the more fine-grained adverb categories.

With respect to lexicon updates concerning nouns and adjectives, table 3 shows that the most frequent type in this study involves correcting morphological properties of nouns concerning the distinction between mass terms and countables. Further, the data indicate that also for nouns and adjectives there is a considerable need for adding subcategorization frames involving MWEs.

Multiword expressions (MWEs) present a great challenge for parsing because they exceed word boundaries, have unpredictable morphosyntactic properties and are sometimes discontiguous (i.e., other words and constituents may come between their component words in a sentence). Treating them as simplex words will thus often result in incorrect or missing analyses. The most immediate problem with MWEs simply concerns knowing about them (Losnegaard et al., 2012), and although there are a considerable number of MWE entries in NorKompLeks (more than 2500 prepositional verbs, 1800 particle verbs and almost 400 fixed expressions), these are not sufficient to account for all of the MWEs in our corpus. For instance, both verbs, nouns and adjectives may take prepositional arguments, while NorKompLeks only provides this kind of subcategorization frame for verbs.

Such frames are added to the lexicon by augmenting the relevant predicates with a preposition or an adverb. Examples of such additions are *legge ut* 'pay', *mening med* 'point of', and *opptatt med* 'concerned with' (examples 8, 9 and 10), where new predicates have been added to the entries for the verb *legge*, the noun *mening*, and the adjective *opptatt*, respectively.

(8) *Jeg måtte legge ut for deg, for jeg*
I must.pret lay out for you, since I
regnet med at du ville gi ham
counted with that you would give him
tips, ikke sant?
tip.pl, not true?
'I had to pay for you since I reckoned you wanted to tip him, right?'

(9) *Hva var da meningen med å sette*
What was then meaning.def with to put
meg i slik forlegenhet?
me in such embarrassment?
'What was the point of embarrassing me like that?'

(10) *Hun ble veldig opptatt med å børste*
she became very busy with to brush
kakesmuler av kåpa si.
cake crumbs off coat.the refl
'She became very concerned with brushing cake crumbs off of her coat.'

Other types of MWE frames that have been added to the lexicon during parsebanking are fixed expressions and verbal idioms. Fixed MWEs are invariable expressions that do not have a normal syntactic buildup. It is thus the expression as a whole, and not the individual words, that must be assigned a lexical category. An example of a *fixed* MWE is *på tå hev* 'on one's toes'. Since *på tå hev* is a completely invariable prepositional phrase, it is added to the lexicon as a word-with-spaces entry, i.e. it appears in the c-structure as one node, as if it were a single word. Because of their syntactic properties, such prepositional phrases with predicative function are classified as adjectives in NorGram.

The addition of words-with-spaces to the lexicon during parsebanking results in a coherently classified inventory of fixed MWEs. In the present study, there were twelve fixed MWEs added; ten updates involved new adverbs, and the other two produced a new adjective entry and a new preposition entry.

While adding lexical entries for hitherto unanalyzed MWEs is an important factor for increasing parsing coverage, there are other and perhaps more general problems associated with the automatic analysis of multiword units. Conventional dictionaries usually provide limited information about MWEs, and their treatment is sometimes incomplete or incoherent. One problem is that the expressions are often not given lexical entries, but only used as examples in the definitions of single-word entries. This information is difficult to extract when constructing an electronic lexicon. For example, in *Bokmålsordboka*, *på tå hev* occurs as an example both under the entry for *tå* 'toe' and under the entry for *heve* '(to) raise', but it does not occur as an entry of its own. Similarly, the prepositional verb *tenke på* 'think about' is listed as one of two senses under the verb *tenke*, but is not explicitly marked as an idiomatic construction. The same MWE is also found under the entries for *andakt* 'piety', *annen* 'other', *fordel* 'advantage', the lexicalized compound *giftetanker* 'marriage plans', and several other semantically unrelated entries.

The perhaps biggest challenge in parsing MWEs is posed by MWEs with internal syntactic structure, such as verbal idioms. These are variable in the sense that they may undergo inflection and syntactic transformations, but idiosyncratic because they may undergo some, but not all kinds of transformations. Although we have identified them as MWEs, it is therefore not straightforward how they should be treated in the lexicon.

An example of a verbal idiom with metaphorical meaning and a regular syntax is *feie under teppet* 'sweep under the carpet'; the verb *feie* requires both a subject and an object in addition to the obligatory PP *under teppet* 'under carpet.the'. However, many MWEs are irregular: the idiom *komme (noen) i møte* 'come (someone) in meeting' ('approach') is idiosyncratic because the verb *komme* 'come' is normally intransitive. Others MWEs form 'families' of apparently similar surface structures.

(11) *ta/ha/få* ... *tak/grep* *(i/på)*
take/have/get … hold/grip (in/on)
'take a hold of/get a hold of/get a (good) grip on, etc.'

Although these constructions seem similar, a closer investigation shows that they have several different (although often related) meanings and that they also differ in their possible syntactic variations.

We cannot expect to find this kind of detail of linguistic description of MWEs in a regular dictionary, and the treatment of MWEs in computational dictionaries varies greatly depending on the type of dictionary, the language in question, and the theoretical framework used. In this respect, parsebanking provides a unique method for detecting problematic constructions such as MWEs, and for acquiring more knowledge about them.

7. Conclusion

Correct lexical information is essential for successful syntactic analysis, but lexical resources derived from dictionaries lack much necessary information, because they are typically not tested in parsing. In our experience, parsebanking is therefore a useful and necessary context not only for grammar development, but also for lexicon development. The INESS project is building up a richer lexical resource for Norwegian and will continue to do so during the remainder of the project. The resulting reusable lexical resource will be made available upon completion of the INESS project in 2016.

8. Acknowledgments

The work reported on here has been supported by the Research Council of Norway and the University of Bergen.

9. References

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

Butt, Miriam, Dyvik, Helge, King, Tracy Holloway, Masuichi, Hiroshi, and Rohrer, Christian. (2002). The Parallel Grammar project. In Carroll, John, Oostdijk, Nelleke, and Sutcliffe, Richard, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics (COLING), Taipei, Taiwan*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dalrymple, Mary. (2001). *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press, San Diego, CA.

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

Hovdenak, Marit, Killingbergtrø, Laurits, Lauvhjell, Arne, Nordlie, Sigurd, Rommetveit, Magne, and Worren, Dagfinn, editors. (1986). *Nynorskordboka : definisjons- og rettskrivingsordbok*. Det norske samlaget, Oslo.

Landrø, Marit Ingebjørg and Wangensteen, Boye, editors. (1993). *Bokmålsordboka: definisjons- og rettskrivnings-ordbok*. Universitetsforlaget, Oslo.

Losnegaard, Gyri Smørdal, Lyse, Gunn Inger, Thunes, Martha, Rosén, Victoria, De Smedt, Koenraad, Dyvik, Helge, and Meurer, Paul. (2012). What we have learned from Sofie: Extending lexical and grammatical coverage in an LFG parsebank. In Hajič, Jan, De Smedt, Koenraad, Tadić, Marko, and Branco, António, editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 69–76, Istanbul, Turkey.

Maxwell, John and Kaplan, Ronald M. (1993). The interface between phrasal and functional constraints. *Computational Linguistics*, 19(4):571–589.

Nordgård, Torbjørn. (2000). Nordkompleks – A Norwegian computational lexicon. In *COMLEX 2000 Workshop on Computational Lexicography and Multimedia*

Dictionaries, pages 89–92, Patras, Greece. University of Patras.

Rosén, Victoria, Meurer, Paul, and De Smedt, Koenraad. (2007). Designing and implementing discriminants for LFG grammars. In King, Tracy Holloway and Butt, Miriam, editors, *The Proceedings of the LFG '07 Conference*, pages 397–417. CSLI Publications, Stanford.

Rosén, Victoria, Meurer, Paul, and De Smedt, Koenraad. (2009). LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Van Eynde, Frank, Frank, Anette, van Noord, Gertjan, and De Smedt, Koenraad, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht. LOT.

Rosén, Victoria, De Smedt, Koenraad, Meurer, Paul, and Dyvik, Helge. (2012a). An open infrastructure for advanced treebanking. In Hajič, Jan, De Smedt, Koenraad, Tadić, Marko, and Branco, António, editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

Rosén, Victoria, Meurer, Paul, Losnegaard, Gyri Smørdal, Lyse, Gunn Inger, De Smedt, Koenraad, Thunes, Martha, and Dyvik, Helge. (2012b). An integrated web-based treebank annotation system. In Hendrickx, Iris, Kübler, Sandra, and Simov, Kiril, editors, *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11)*, pages 157–167, Lisbon, Portugal. Edições Colibri.