What We Have Learned from Sofie: Extending Lexical and Grammatical Coverage in an LFG Parsebank

Gyri Smørdal Losnegaard*, Gunn Inger Lyse*, Martha Thunes*, Victoria Rosén*§, Koenraad De Smedt*, Helge Dyvik*§, Paul Meurer§

University of Bergen* and Uni Research§
Bergen, Norway
gyri.losnegaard@lle.uib.no, gunn.lyse@lle.uib.no, martha.thunes@lle.uib.no, victoria@uib.no, desmedt@uib.no, paul.meurer@uni.no, dyvik@uib.no

Abstract

Constructing a treebank as a dynamically parsed corpus is an iterative process which may effectively lead to improvements of the grammar and lexicon. We show this from our experiences with semiautomatic disambiguation of a Norwegian LFG parsebank. The main types of grammar and lexicon changes necessary for achieving improved coverage are analyzed and discussed. We show that an important contributing factor to missing coverage is missing multiword expressions in the lexicon.

1. Introduction

The INESS project[1] (2010–2015) is building a highly detailed treebank for Norwegian by parsing corpora with the NorGram LFG grammar (Dyvik, 2000; Butt et al., 2002). A major challenge for the automatic analysis of corpora is incomplete coverage by the grammar and lexicon, in particular related to the phenomenon of multiword expressions. These are poorly documented in most languages, including Norwegian. Although NorGram has an extensive lexicon, few multiword expressions are included. In our work with semiautomatic disambiguation of corpora we have noted that multiword expressions missing from the lexicon cause many gaps in coverage, a problem that often may be solved by simply adding the relevant expression to the lexicon.

In this paper we present a study of a small subcorpus in the INESS Norwegian treebank. This study diagnoses some of the reasons for missing coverage, and we show that a large proportion of the problems are caused by missing multiword expressions.

2. The parsebanking method in INESS

The parsebanking method used in the INESS project involves parsing, disambiguation, and grammar and lexicon development in an iterative cycle. This method was perhaps first developed for HPSG grammars in the LinGO project (Oepen et al., 2004). Overviews which describe this type of approach are found in Branco (2009) and Bender et al. (2011).

In our approach, a corpus is first parsed automatically using the Xerox Linguistic Environment (XLE) (Maxwell and Kaplan, 1993) and the NorGram LFG grammar. LFG analyses provide two separate but parallel levels of syntactic analysis. There is a constituent structure (c-structure) in the form of a context-free phrase structure tree, and a functional structure (f-structure), an attribute–value matrix with information on grammatical features and syntactic functions.

Since automatic parsing with a handwritten grammar produces many analyses, efficient disambiguation is necessary. This is done using discriminants, as described in more detail elsewhere (Carter, 1997; Oepen et al., 2004; Rosén et al., 2007; Rosén et al., 2009). After disambiguation has been achieved, it becomes apparent whether the intended analysis is present. When this is not the case, the annotators try to diagnose the reason for the problem. This may be done by experimenting with XLE-Web (Rosén et al., 2005), for instance by changing or deleting words or phrases to check if modified sentences get an analysis, and suggesting changes to the grammar or lexicon. After necessary changes have been made in the grammar and lexicon, the corpus is reparsed.

After each reparsing, the corpus is automatically disambiguated by means of the reapplication of cached discriminants. In some cases the previously used discriminant choices are no longer sufficient to fully disambiguate the sentence due to the changes made in the grammar and lexicon. In such cases, some additional discriminant choices must be made by the annotators.

A major advantage of this approach is that the analyses in the treebank are always in accordance with the grammar, and coverage may be improved in a principled way.

3. Study of the Norwegian Sofie treebank

For this study we have investigated the first 255 sentences of the novel *Sofies verden* [Sophie's World] (Gaarder, 1991) to find out what the major problem types are that need to be addressed in order to achieve coverage of a subcorpus. The 255 sentences were initially parsed without prior examination of their vocabulary. Then followed several rounds of disambiguation, problem diagnosis, grammar and lexicon changes, and reparsing. The results of parsing in the first and fourth rounds are shown in Table 1 (all numbers are percentages).

version	gold	not gold	frag	0 sol	no parse
1	26	21	26	5	22
4	78	2	4	14	2

Table 1: Initial and subsequent parse results

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

In the first round, 73% of the sentences received analyses (the categories ‘gold’, ‘not gold’ and ‘frag(ment)’, while 27% received no analyses (the categories ‘0 sol(utions)’ and ‘no parse’).

There were full analyses for 47% of the sentences, while 26% of the sentences had the intended analysis (gold) without any intervention in the grammar or lexicon.

Fragment analyses occur when the grammar is unable to assign a global analysis to the sentence, and instead returns the analyses of the maximal phrases which it has been able to parse. Fragment analyses may indicate shortcomings of the grammar, or they may indicate strings deviating from the grammatical norms of the language (Rosén and De Smedt, 2007). In the latter case the fragment analyses are considered as valuable information worth storing. An example from the corpus is the sentence - *Det ... det er en hemmelighet* ‘It ... it is a secret’, with the dots signaling hesitation on the part of the speaker. The fragment analysis is shown in Figure 1, indicating the analyzable chunks of the string.

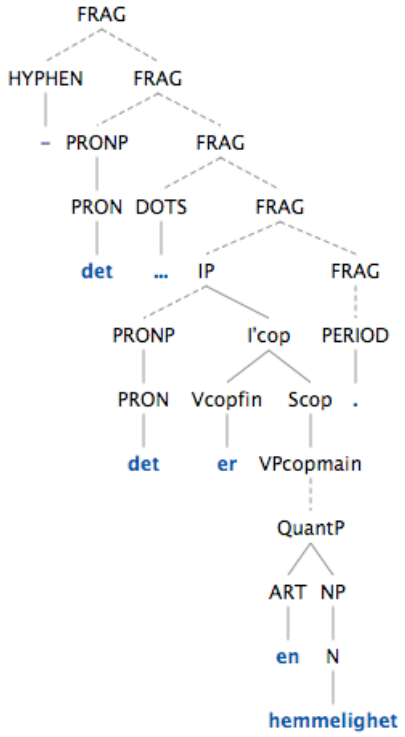

Figure 1: Fragment analysis

The category ‘0 solutions’ means that the parser terminated, but found no analysis in accordance with the grammar, while the category ‘no parse’ means that the parser didn’t terminate within the time and space parameters set for it. As we will show, the first category was significantly reduced as a result of the interaction between annotators and grammar developer, while the second category grew as a result of increased grammar complexity, leading to occasional explosions in local ambiguities. The latter problem will be addressed later in the project by devising methods for shallow preprocessing of sentences, thus reducing local ambiguity.

4. Problem analysis

We have done an in-depth study of the parsed sentences that were missing the intended analysis in version 1 but that did receive the intended analysis in version 4. In particular, we have studied the interventions that were necessary in order to produce the desired analysis. We have concentrated on the sentences that originally had full analyses rather than fragments, since these have received the most attention so far in disambiguation and problem diagnosis. For a small number of these sentences (2%) it was not possible to ascertain why they did not get the intended analysis originally. This could be because the problem diagnosis was not recorded. For most of the sentences, we have, however, identified the problem or problems; some sentences had multiple problems. For this set of sentences (the 21% ‘not gold’ in version 1) the problems may be analyzed into two main categories: grammar problems (29%) and lexicon problems (71%). The two most prevalent types of lexicon problems are multiword expressions and lexical category problems, which make up 41% and 31% of the lexicon problems respectively. In the following sections we discuss the various problem types.

4.1. Grammar problems

Under the category of grammar problems we have considered various instances of shortcomings in the rule component of the grammar. In these cases it is necessary to extend the grammar by including types of constructions that have not yet been covered, and in order to solve such problems, the grammar writer must face the challenge of describing exactly the necessary classes of constructions, while avoiding the introduction of changes that may cause the grammar to overgenerate. Required changes may involve the writing of new phrase structure rules, as well as the modification of existing rules. Reported grammar problems may be illustrated by two examples.

(1) *Et menneske måtte da være noe mer*
a human must then be something more
enn en maskin?
than a machine
‘A human then had to be something more than a machine?’

In (1) the original analysis of the comparative construction *noe mer enn en maskin* was not satisfactory. The quantifier *noe* was recognized as an adverb of degree (ADVdeg), giving the meaning ‘somewhat more’, and the expression *noe mer* was parsed as a quantifier phrase (QP), as shown in the c-structure in Figure 2. The intended analysis was achieved by modifying the rule describing quantifier phrases, and the modification involved allowing for recursivity in phrases of this category. In simplified terms, recursivity was introduced on the condition that the left-most Q must be a form of the quantifier *noe*. This resulted in a new analysis of the phrase *noe mer enn en maskin* where *noe* is correctly parsed as a quantifier, and the entire expression is assigned the hierarchical structure shown in Figure 3.

(2) *Uansett hva Sofie gjorde, gjorde hun*
regardless what Sofie did did she
akkurat det samme.
exactly the same
‘No matter what Sofie did, she did just the same.’

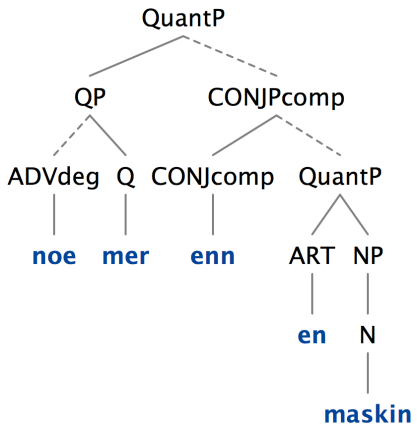

Figure 2: Incorrect analysis of *noe mer*

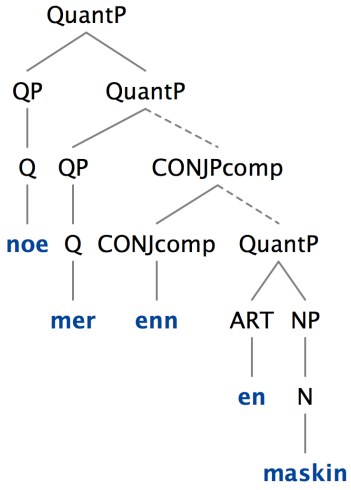

Figure 3: Correct analysis of *noe mer*

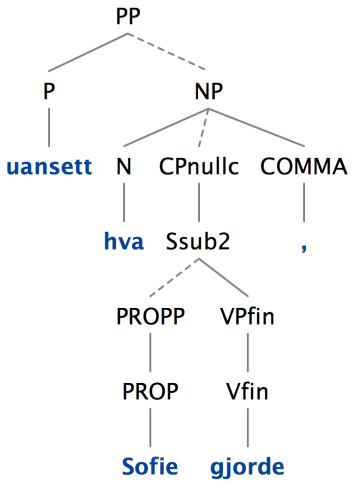

Figure 4: Incorrect analysis of *hva Sofie gjorde*

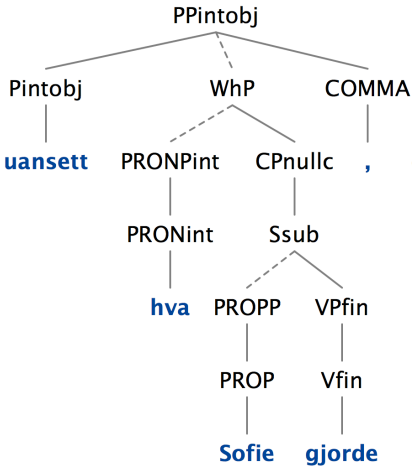

Figure 5: Correct analysis of *hva Sofie gjorde*

In (2) an unsatisfactory analysis of the expression *uansett hva Sofie gjorde* was amended by introducing a new phrase structure rule allowing certain prepositions to take interrogative constructions as their object. Originally, the expression *hva Sofie gjorde* was wrongly recognized as a noun phrase with a clausal postmodifier, and in that analysis the phrase was treated as the argument of the preposition *uansett*, as shown in Figure 4. Here the grammar update involved creating a special rule for prepositional phrases taking interrogative phrases as object (PPintobj), as shown in Figure 5, and the application of this rule is restricted to a limited set of prepositions, of which *uansett* is an example. Solving this problem also involved adding to the lexicon a new reading of *uansett* with the lexical category Pintobj. In the new analysis we achieved a satisfactory parse of the interrogative expression *hva Sofie gjorde*, where *hva* is correctly recognized as an interrogative pronoun.

In NorGram the analysis of punctuation is incorporated in the rule component of the grammar. Hence, cases where sentences have not been parsed because the parser has not recognized specific uses of punctuation marks may in the context of the present study be regarded as special cases of grammar problems. Among the challenges encountered in automatic parsing of authentic text is the correct analysis of various kinds of punctuation marks. Since the orthographic conventions governing the use of punctuation marks are not always very clear or generally agreed upon, it is not a trivial task to handle the various possible uses of different marks. In the Sofie treebank we have observed several cases where sentences have not been parsed successfully because particular ways of using specific punctuation marks were not covered by the grammar rules. Some examples of the use of dashes may illustrate this type of challenge.

(3) *Der lå et prospektkort også — med bilde av en sydlig strand*
there lay a postcard also — with picture of a southerly beach
'There was a postcard too — with a photo of a southern beach.'

(4) *Joda — det var ekte nok, med både frimerke og stempel.*
yes it was genuine enough with both stamp and postmark
'Oh yes — it was real enough, with both a stamp and a postmark.'

(5) *Sofie stirret ned i asfalten — og opp på venninnen.*
Sofie stared down in asphalt:the and up on girlfriend:the
'Sofie stared down at the asphalt — and back at her friend.'

In these three sentences the effect of the dash is to create a short pause, and in the cases of (4) and (5) this pause puts a slight emphasis on the expressions following the dashes. In the original treebank version the parser produced a fragment analysis for each of these sentences. With respect to (3), this was the case because the dash prevented the parser from recognizing the prepositional phrase *med bilde av en sydlig strand* as being part of the sentence, and in the case of (5), the initial word *joda* was not incorporated in the sentence because of the following dash. As regards sentence (5), it contains two coordinated PPs, *ned i asfalten* and *opp på venninnen*, which are linked together by the conjunction *og*, but the parser fails to recognize the coordination because of the dash preceding the conjunction. In each of these cases, substituting a comma for the dash gave a full analysis of the sentence, and the problems were solved by amending the rule component of the grammar to allow dashes on a par with commas in these types of syntactic positions.
These three examples illustrate a few points that are common to several cases of punctuation problems. Firstly, to solve such problems it may be necessary to modify several grammar rules; in relation to the dash, rules applying to sentence as well as to verb phrase level were involved. Secondly, because the use of several types of punctuation marks, such as dashes, colons, and quotes, may be rather idiosyncratic, i.e. governed by individual authors' preferences, fairly ad hoc solutions may be required by the grammar developer in order to account for the various uses of different punctuation marks. The practical consequence of the latter point is that the punctuation found in a given corpus may to a great extent determine the ways in which the grammar writer chooses to solve the punctuation problems at hand.

4.2. Problems with multiword expressions

The largest group of lexicon problems encountered in our analysis had to do with multiword expressions (MWEs), which can roughly be defined as "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002). MWEs challenge the division between grammar and lexicon in linguistic theory due to the fact that they are lexicalized, but they may show variation at the morphological, syntactic and semantic levels.
Varying in terms of syntactic flexibility, MWEs can be grouped into the subcategories *fixed*, *semi-fixed* and *syntactically-flexible* expressions (Baldwin and Su Nam Kim, 2010; Sag et al., 2002). Semi-fixed and syntactically flexible expressions pose the biggest challenges in automatic analysis because they inflect, take internal modification or in other ways realize morphosyntactic variation. Different morphosyntactic categories of MWEs tend to adhere to different flexibility categories; for instance, verbal constructions are generally syntactically flexible. Depending on their flexibility, MWEs can be treated as words with spaces (fixed expressions) or as constructions. Constructions may either deviate from the morphosyntactic regularities of the language or be fully compositional in the sense that the rules of the grammar capture their syntactic and semantic properties satisfactorily. In these cases we have left them 'as they are' and not explicitly marked them as MWEs.
Sag et al. (2002) single out four main problems related to the representation of MWEs in NLP systems. Treating MWEs as a problem of the lexicon poses a *flexibility problem* because we fail to capture morphosyntactic flexibility such as internal modification, or cases where some constituents inflect while others do not. Listing every single MWE in the lexicon also leads to a *lexical proliferation problem*: we lose out on generalities such as families of verbal constructions. By treating MWEs as a problem of grammar, the application of general compositional methods will lead to an *overgeneration problem* because the grammar will allow for anti-collocations and other unacceptable constructions. Finally, *idiomaticity problems* may occur because grammar rules, working on sentence level, cannot distinguish between literal and figurative meanings.
In their diagnosis, the annotators reported all instances of possible MWEs that might have caused problems. The decision on which expressions should be implemented as MWEs and which should get a compositional analysis was made for each instance. With lexicon overpopulation in mind, we have as far as possible tried to analyze light verb constructions compositionally, although their deviating semantics clearly indicate what Baldwin and Kim refer to as MWEhood (Baldwin and Su Nam Kim, 2010). We cannot be certain that similar expressions have always been analyzed the same way, but as we gain experience in this early phase of annotation and get a better overview of the different types of MWEs, we will eventually be able to have a more principled approach to their identification and implementation.
For this study we have chosen to classify MWEs as a lexicon problem because of the practical implications of their analysis within the LFG framework: all implementations of MWEs have so far taken place in the lexicon, and not at rule level. Their syntactic variation is still fully accounted for by the grammar.
We have distinguished between two types of MWEs: verb frames and other MWEs. MWE verb frames include light verb constructions (*ta slutt* 'end'), particle verbs (*se ut* 'look'), and selected prepositions (*vite om* 'know about', 'be aware of'). Light verb constructions are verbs with noun complements where the verb meaning has become semantically 'light' compared to the contribution of the noun to the meaning of the overall expression (Baldwin and Su Nam Kim, 2010). Particle and prepositional verbal constructions have one or more associated lexical items which modifiy the verb predicate, making the compositional analysis (literal meaning) unacceptable. As a particle verb, *se ut* means 'look' in the sense 'look like' or 'appear', while its literal meaning is 'look out', as in *se ut gjennom vinduet* 'look out (through) the window'.
Verbal constructions have been implemented as verbs with non-thematic objects (light verb constructions), as particle

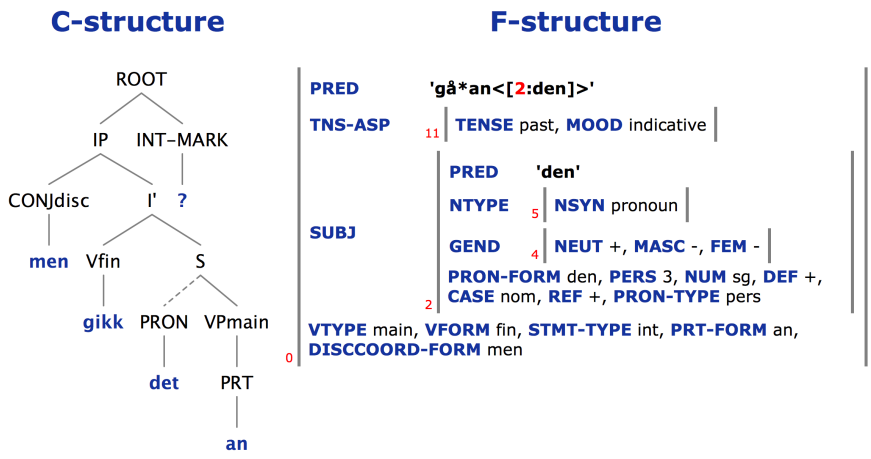

Figure 6: Constituent and functional structure representations with particle verb

verbs or as verbs with selected prepositions, depending on their syntactic properties. The LFG distinction between c- and f-structure allows NorGram to capture the compositional and non-compositional properties of MWEs in a perspicuous way. Thus, all the verb frame MWE types mentioned above express the non-compositional meaning of the MWE as augmented predicates in the predicate-argument structure: 'ta*slutt<(↑SUBJ)>(↑OBJ)', 'se*ut<(↑SUBJ)>', 'vite*om<(↑SUBJ)(↑OBL-TH)>'. The words *slutt*, *ut* and *om* are analyzed respectively as a non-thematic object (outside the argument frame) selected by the verb entry, a selected particle, and a selected preposition. The c-structure captures the regular syntactically productive analysis of the expressions, allowing the usual variations in constituent order. Only when the MWEs allow no (or insignificant) formal variation and no intervening words may they be analyzed as 'words with spaces' in the lexicon.
An example of a particle verb from the Sofie corpus is provided in 6; the particle *an* is only found in MWEs.

(6) *Men det gikk an.*
but it went PRT
'But it was possible.'

The MWEhood of the particle verb construction is shown by the complex predicate name in the value of the PRED feature in the f-structure in Figure 6. Examples of 'words with spaces' are expressions such as *med ett* 'suddenly' and *borte vekk* 'gone', which are fixed and thus simply added to the lexicon with the appropriate lexical category.
We also recorded problems with MWEs that belong to the group of semi-fixed expressions. As an example, the annotators suggested that the multiword unit *et eller annet* 'some, something' (literally 'one or another') was the main problem in the unsatisfactory analyses of the sentences in 7 and 8.

(7) *Altså måtte verdensrommet en eller annen*
thus must space:the one or another
gang ha blitt til av noe annet.
time have become to of something else
'So space must once have been created from something else.'

(8) *Til syvende og sist måtte et eller annet*
to seventh and last must one or other
en gang ha blitt til av null og
one time have become to of zero and
niks.
nought
'Ultimately something must once have been created from diddly-squat.'

The expression was added to the lexicon as a MWE quantifier (Q). Like many quantifiers in Norwegian, the MWE agrees with the quantified nominal, and two of its components — the determiner *en* and the adjectival determiner *annen* — inflect in gender. This was implemented by adding three different entries to the lexicon, one for each gender (*en eller annen* M, *ei eller anna F*, *et eller annet* NEUT). Adding one entry in the lexicon for each possible form of a MWE is not very economical. If we adopt this practice we risk overpopulation, and we fail to capture the morphological generalities of the set of inflectional forms, both common problems when treating MWEs as problems of the lexicon as described by Sag et al. (2002). These kinds of problems thus call for being somewhat restrictive with respect to adding new MWEs to the lexicon, and how to implement them. We want to capture as many MWEs as possible, but we also want to avoid representing every exception to the grammar rules in the lexicon. However, such constructions are so prevalent that they pose problems for syntactic analysis and thus cannot be ignored, amply illustrated by the sentence in 8. In the phrase structure tree representing its constituent structure, shown in Figure 7, the number of terminal nodes is far smaller than the number of individual words, reflecting the fact that there are four MWEs in the sentence.
MWE implementation accounts for 41% of the lexicon updates carried out after the first parse of Sofie. The main question remains how to represent flexibility, and how to account for families of constructions and other kinds of structural frames. Given the frequency of MWEs among the problems encountered, it seems reasonable to assume that the further grammar development would benefit from applying fixed criteria for MWEhood and from identifying

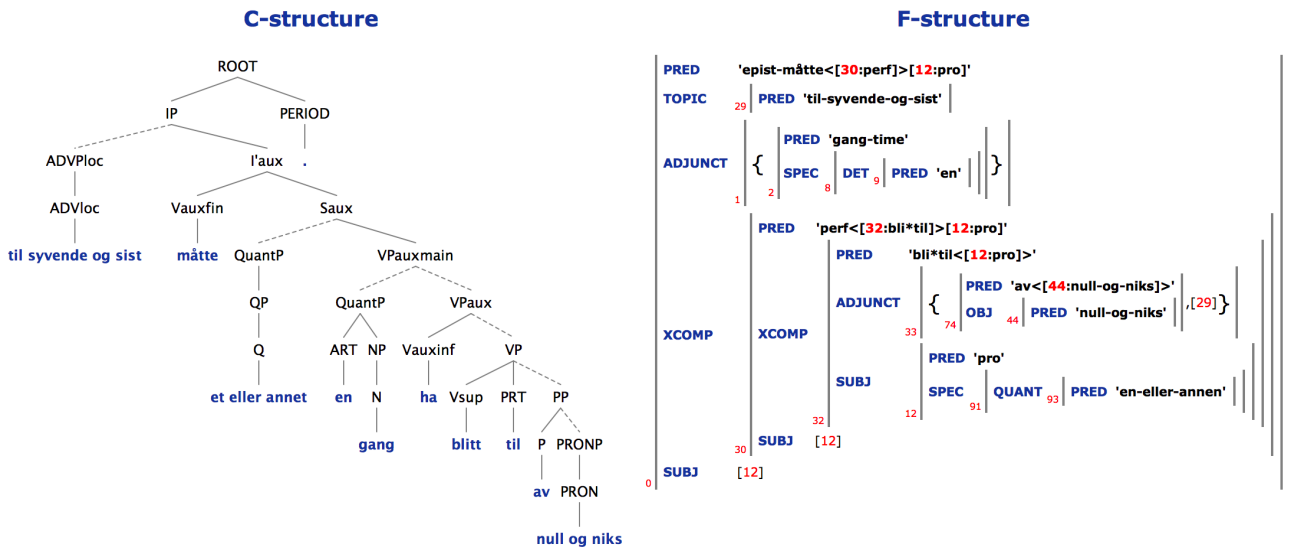

Figure 7: Sentence with multiple MWEs

optimal techniques of analysis for the implementation of different types of MWEs. This could provide a more consistent treatment of MWEs in the grammar, and make us better equipped to meet the challenges MWEs pose in automatic processing.

4.3. Other lexicon problems

Following MWEs, the major group among the recorded lexicon updates is lexical category updates, with 31 % of the reported problems. In addition, we found some problems related to lexical frames (or subcategorization), new lexicon entries, and new word senses added to the lexicon. Lexical category updates are cases where words were reclassified after a problem had been reported for the sentences in which they occurred. An example is *riktignok* 'true', 'indeed', which was previously classified as a verb phrase adverb (ADV) in the lexicon.

(9) *Et menasjeri var en samling av forskjellige*
a menagerie was a collection of different
dyr, og riktignok — Sofie var ganske
animals and indeed Sofie was quite
godt fornøyd med sin egen samling.
well content with her own collection
'A menagerie was a collection of different animals, and Sofie was indeed quite content with her own collection.'

The part of speech *adverb* is a large class which encompasses many words with quite different syntactic properties. The words traditionally described as adverbs have verb phrase adverb as their default classification in our lexicon. In the case of *riktignok*, it was the syntactic distribution of this category, as specified in the grammar rules, that was found to cause problems for the constituency analysis. The example illustrates that lexical category updates are often grammatically motivated in the sense that words with a certain classification in the lexicon sometimes turn out to have syntactic properties that single them out as a separate category. After annotator intervention, *riktignok* was additionally classified as a root adverb (ADVroot). Root adverbs like *riktignok* differ from other adverbs because they can form utterances on their own.

The lexical categories in NorGram are fundamentally based on syntactic distribution, and by parsing Sofie we discovered several instances of previously unseen syntactic behavior of different lexical categories, or members of these categories, leading to reclassification in the lexicon. This may be further exemplified by the class of reflexive pronouns, which after intervention is divided into two categories, non-referring and referring reflexives. A referring reflexive in Norwegian is the MWE *seg selv* in sentences such as *barna vasker seg selv* 'the children wash themselves', as opposed to the non-referring *seg* in *barna vasker seg* 'the children wash', where *seg* is used with the reflexive verb *vaske*. Previously, NorGram did not have an analysis for the special case of referring *seg*, as found in 10 and 11.

(10) *Straks Sofie hadde lukket porten*
immediately Sofie had closed gate:the
bak seg, åpnet hun konvolutten.
behind self opened she envelope:the
'As soon as Sofie had closed the gate behind her, she opened the envelope.'

(11) *Sofie skyflet katten ut på trappen og*
Sofie shoved cat:the out on stair:the and
lukket døren etter seg.
closed door:the after self
'Sofie shoved the cat out onto the stairs and closed the door behind her.'

The examples show that the simple form *seg* can also be a referring reflexive in certain contexts, and due to these findings, this case is now singled out as a special category, PRONrfl2, restricted by the grammar to the relevant contexts.

The next type of lexicon update is the one we termed 'lexical frames'. These may also be defined as grammatically

motivated updates, for instance when a word does not have the needed valency. An example is the verb *huske* 'remember', which was found to lack an intransitive reading. As a result, intransitive *huske-remember* has now been added to the lexicon. We also encountered an unacceptable analysis of a sentence with the MWE candidate *komme rekende på en fjøl*.

(12) *Det hadde bare kommet rekende på en fjøl.*
it had only come drifting on a board
'It had just appeared from nowhere.'

Instead of treating the whole expression as a MWE, we accounted for the construction *komme rekende*, which is an instance of a general infinitive–present participle construction restricted to certain verbs, and which was already implemented. The necessary update was made in the lexical entry for the verb *reke*.

As a result of the problem diagnostics, four new interjection-like words were also added to the lexicon as root adverbs (ADVroot). These were the words *næh* 'nah', *neivel* 'no indeed', *pøh* 'bah', and *fillern* 'darn'. Like most interjections, these are all very colloquial with a fairly unorthodox orthography.

The final type of lexicon update is the addition of new senses to words that are already in the lexicon. We recorded two cases of missing senses, of which two concerned the same word, the noun *gang* 'corridor'. One instance was found in the sentence *I neste øyeblikk var hun ute i gangen* 'The next moment she was out in the hallway'. The main problem was actually not the sense in itself, but restrictions on semantic features associated with that particular sense, since the grammar sometimes requires that semantic features must be checked against features of associated words. During the first parse, *gang* was only implemented as a temporal noun in the lexicon. After update, the lexicon now distinguishes between *gang-time* and *gang-corridor*. This also allows using the 'regular', non-temporal preposition *i* 'in' together with *gang*, something which was not possible after the first parse and which was the direct cause of the unintended initial analysis. Since the only implemented analysis of *gang* was temporal, only *i* with temporal features could be used with this noun. Having added the non-temporal *gang-corridor*, we are now able to choose the non-temporal *i*.

Among the problems identified as lexical updates, many proved to be problems also of the grammar. This demonstrates that the different components of a (computational) grammar are closely interrelated, and that updates to one of the components will almost certainly affect the other. We have several times experienced unfortunate outcomes of grammar and lexicon updates, and keep in mind that any change to the grammar bears a risk of allowing for overgeneralization and other unwanted side-effects.

4.4. Related work

Some related research has been aimed at analyzing causes of missing parser coverage and improving coverage by automatic lexical acquisition. Nicholson et al. (2008) give a breakdown of gaps in coverage, including 7% due to missing MWEs, and propose the addition of lexical entries by hypothesizing their types. In a similar vein, Villavicencio et al. (2007) cite 8% parse failures due to missing MWEs and propose a lexical type predictor as well. Other lexical type predictors are proposed by Cholakov and van Noord (2010), who assign lexical types to single words only, and Zhang et al. (2010), who take a step towards acquiring MWEs, although with limits on the valency and complexity of the covered constructions. Some of these methods seem compatible with our parse methods, but so far it seems that accurate remedial actions for MWEs still need a manual intervention step.

5. Conclusion

Treebanking by the automatic parsing of corpora is a way of validating a computational grammar and lexicon, thereby identifying gaps in the computational treatment of a language. The lexicon for Norwegian used in INESS is quite extensive and in general provides excellent coverage. Existing resources for Norwegian — and also for other languages — are, however, quite lacking with respect to MWEs. The lexical resources for Norwegian contain a number of MWEs which were taken from traditional dictionaries or from NorKompLeks (Nordgård, 1998), but the latter was also based on dictionary examples, not on corpus study. Our experience shows that these are insufficient. Although our small study showed that there were two main types of lexicon problems, lexical category problems and problems with multiword expressions, these two categories differ in an essential way. The lexical category problems concern the category assigned to items already within the lexicon. Solving these problems involves changing the lexical category. The big challenge posed by multiword expressions is that we do not have access to an inventory over them. Solving these problems is therefore an undertaking of a much greater dimension.

The interactive and iterative approach adopted in INESS makes it possible to integrate more and more MWEs into the lexicon and grammar and to thereby reach better coverage, especially in the semiautomatically disambiguated part of the treebank described in this paper. We will, however, also create a large parsebank that will be fully automatically disambiguated, and for this, we cannot rely on discovering MWEs during the annotation process.

The work done with lexical issues in INESS will contribute to improving the lexical resources available for Norwegian. But it is important that independent work is done on finding and analyzing MWEs, not only for Norwegian but also for other languages.

6. Acknowledgments

This research has received funding from the Research Council of Norway under the program National Financing Initiative for Research Infrastructure.

7. References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, chapter 12. CRC Press, Boca Raton, FL, USA, second edition.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2011. Grammar engineering and linguistic hypothesis testing: Computational support for complexity in syntactic analysis. In Emily M. Bender and Jennifer E. Arnold, editors, *Language from a Cognitive Perspective: Grammar, Usage, and Processing*, CSLI Lecture Notes 201, pages 5–29. CSLI Publications.

António Branco. 2009. LogicalFormBanks, the next generation of semantically annotated corpora: key issues in construction methodology. In Mieczyslaw Klopotek, Adam Przepiórkowski, Slawomir Wierzchón, and Krzysztof Trojanowski, editors, *Recent Advances in Intelligent Information Systems*, pages 3–12. Academic Publishing House EXIT, Warsaw.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan*.

David Carter. 1997. The TreeBanker: A tool for supervised training of parsed corpora. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 598–603, Providence, Rhode Island.

Kostadin Cholakov and Gertjan van Noord. 2010. Acquisition of unknown word paradigms for large-scale grammars. In Chu-Ren Huang and Dan Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics, Beijing*, volume Posters of *COLING '10*, pages 153–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

Helge Dyvik. 2000. Nødvendige noder i norsk. Grunntrekk i en leksikalsk-funksjonell beskrivelse av norsk syntaks [Necessary nodes in Norwegian. Basic features of a lexical-functional description of Norwegian syntax]. In Øivin Andersen, Kjersti Fløttum, and Torodd Kinn, editors, *Menneske, språk og felleskap*. Novus forlag.

Jostein Gaarder. 1991. *Sofies verden: roman om filosofiens historie*. Aschehoug, Oslo, Norway.

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

Jeremy Nicholson, Valia Kordoni, Yi Zhang, Timothy Baldwin, and Rebecca Dridan. 2008. Evaluating and extending the coverage of HPSG grammars: A case study for German. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Torbjørn Nordgård. 1998. Norwegian computational lexicon (NorKompLeks). In *Proceedings of the 11th Nordic Conference on Computational Linguistics (NoDaLiDa), Copenhagen*.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christopher D. Manning. 2004. LinGO Redwoods, a rich and dynamic treebank for HPSG. *Research on Language & Computation*, 2(4):575–596, December.

Victoria Rosén and Koenraad De Smedt. 2007. Theoretically motivated treebank coverage. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NoDaLiDa-2007)*, pages 152–159. Tartu University Library, Tartu.

Victoria Rosén, Paul Meurer, and Koenraad De Smedt. 2005. Constructing a parsed corpus with a large LFG grammar. In *Proceedings of LFG'05*, pages 371–387. CSLI Publications.

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

Victoria Rosén, Paul Meurer, and Koenraad De Smedt. 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, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht. LOT.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Lecture Notes In Computer Science. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2276, pages 189–206. Springer.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043, Prague, Czech Republic, June. Association for Computational Linguistics.

Yi Zhang, Timothy Baldwin, Valia Kordoni, David Martinez, and Jeremy Nicholson. 2010. Chart mining-based lexical acquisition with precision grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 10–18, Los Angeles, California, June. Association for Computational Linguistics.