Multiwords, Word Senses and Multiword Senses in the Eukalyptus Treebank of Written Swedish

Yvonne Adesam, Gerlof Bouma and Richard Johansson

Språkbanken, Department of Swedish
University of Gothenburg
E-mail:{yvonne.adesam|gerlof.bouma|richard.johansson}@gu.se

Abstract

Multiwords reside at the intersection of the lexicon and syntax and in an annotation project, they will affect both levels. In the Eukalyptus treebank of written Swedish, we treat multiwords formally as syntactic objects, which are assigned a lexical type and sense. With the help of a simple dichotomy, analyzed vs unanalyzed multiwords, and the expressiveness of the syntactic annotation formalism employed, we are able to flexibly handle most multiword types and usages.

1 Introduction

The *Eukalyptus treebank of written Swedish* will contain about 100.000 tokens and is under active development. It's foremost purpose is to serve as an evaluation corpus for multiple annotation tools, from part-of-speech taggers over sense disambiguators, to parsers. Because of this, it has from the onset been designed with a range of annotations in mind, which has influenced the design of the individual annotation levels. Previous papers [1, 2] have described the purpose of the project and the syntactic annotation of the treebank. In this paper, we focus on the levels of word senses and syntactic structure, which are connected by the shared concern of multiwords. We show how the issue of multiwords and multiword senses is handled by introducing a simple dichotomy in their syntactic annotation. Because both our syntactic annotators and our word sense annotators are confronted with multiwords, we are also able to give an empirical comparison of their annotations.

2 Annotation Levels

The range of annotations in the Eukalyptus treebank can be summarized as follows. Our token definition is roughly the graphic word. Below the token level, we then annotate compound structure; at the token level, lemmata, word senses, parts of

speech and morphological features; and above, syntactic structure. When dealing with multiwords, above-token-level annotation also includes multiword lemmata, multiword parts of speech and multiword senses.

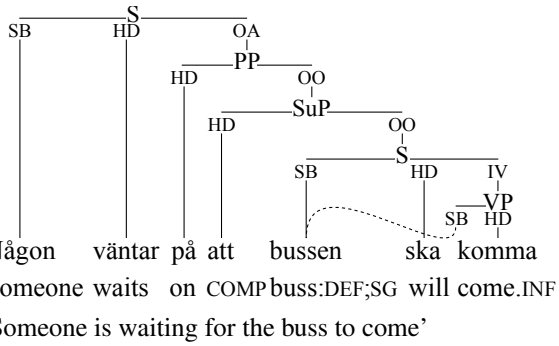
For our inventory of word senses and lemmata, we rely on the SALDO lexical resource [4], which defines senses by placing them in a network of associations. Crucially, SALDO not only contains word senses for single word entries, but, at the time of writing, also for around 8.000 multiword entries, which make up approximately 6.5% of the entries. From the perspective of SALDO, multiword entries are just *word entries*, which means that there is no principled difference in their treatment compared to single word entries. Amongst other things, multiwords are assigned part-of-speech tags in accordance with the regular tag definitions. For instance, there is no concept of 'verb-object idiom' in SALDO, as these are just multiword verbs. For example, the multiword *dra timmerstockar* 'snore' (lit.: 'pull timberloggs') is marked as a multiword verb (VBM), and has *snarka* 'snore' as its primary associative link. Similarly, the expression *lagens långa arm* 'the police' (lit.: 'the law's long arm') is marked as a multiword noun (NNM), with primary link *polis* 'police'. Like SALDO, the Eukalyptus treebank uses parts-of-speech for multiwords. However, in contrast to SALDO and as detailed below, we do, as far as possible, annotate internal syntactic structure in multiwords.

Eukalyptus' syntactic annotation scheme is formally based on the familiar German NEGRA/TIGER scheme [5], combining (possibly discontinuous) phrases with labelled edges for the syntactic functions. A syntactic analysis consists of a primary graph, which is a rooted tree yielding all tokens in the annotation unit, and additional, secondary edges that can be used to express sharing. The combined primary and secondary annotations form an unrestricted directed labelled graph.

We follow, and extend upon, the descriptive traditions of the pioneering annotation guidelines MAMBA [6] from the 1970s and the modern reference grammar *Svenska Akademiens Grammatik* [7]. Phrases in Eukalyptus are generally constrained to be headed by lexical material, and a set of projection rules links the 13 parts-of-speech categories to 10 phrase categories. For each of the 13 parts-of-speech, there is a counterpart multiword part-of-speech, recognizable by a suffix 'M'. However, whereas parts-of-speech formally are terminal node labels in the syntactic tree, multiword parts-of-speech are non-terminal node labels, just like the phrase categories. Non-head children may have one of 20 different grammatical functions, partially depending on the phrase type. An example syntactic tree without any multiwords is given in figure 1.

3 The Analyzed-Unanalyzed Dichotomy: Multiwords as Syntactic Structure

Many types of multiwords have realizations that look like regular syntactic constructions. For instance, a verb-object idiom will take the shape of a non-idiomatic verb object combination, although its variation possibilities may be more or less



Note: Apart from the more common abbreviations, the tree uses phrase label SuP for Subordinator Phrase (similar to CP), and dependency labels OA for bound adverbials, OO for (direkt) objekts/complements, and IV for non-finite verbal complements. The solid lines show the primary tree, the dashed line shows the secondary edge used to indicate the implicit subject of *komma* 'come.INF'.

Figure 1: Syntactic tree for *Någon väntar på att bussen ska komma.*

restricted. From a syntactic annotation perspective, it is attractive to annotate such realizations as regular syntactic structures. The structure may throw light upon some of the regularities we see in the realization, and more importantly, for idioms that allow internal modification, we need the syntactic structure to attach the modifiers in the right place. Consider (1), which involves the multiword *dra timmarstockar* 'snore'.

(1) den andra slutade dra [$_{NP}$ de allra tyngsta timmerstockarna]
the other stopped pull.INF the very heaviest timber logs.DEF
'The other one doesn't snore as heavily as he did before.'

The determiner *de* and adjectival attribute *allra tyngsta* can only attach to *timmerstockar* if the word is actually allowed to head a phrase and is not just considered part of the multiword.

We might therefore consider multiword annotation to be formally independent of syntactic annotation. At some separate level, we would then represent groups of tokens to which we can attach the multiword senses. However, other types of multiwords pose problems for syntax in ways that suggest that multiwords should be represented directly in syntax. For example, the NP in (2) is headed by what looks like a PP. This would not only be unexpected but it would violate Eukalyptus' well-formedness rules on heads, which say that heads be lexical and have a part-of-speech related to the phrasal category.

(2) [$_{NP}$ Anderssons [$_{PP}$ Till min syster]]
Andersson's to my sister
'(Dan) Andersson's (poem) *For my sister*'

However, if we take into account that the 'offending' head is the title of a poem, and can therefore be considered a multiword proper name, we can see that the violation of the well-formedness rules is only apparent: multiword proper names are both lexical and nominal. We can easily adjust our well-formedness criteria to correctly allow (2), if we include information about multiwordhood into the annotation graph.

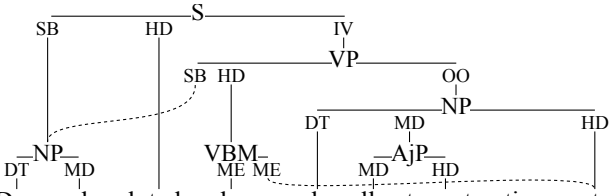
A different problem is found in the multiword proper name in (3), which follows the conventions for person names, but not those for, say, Swedish NPs, without resorting to ad hoc structures. Instead, it appears that it is exactly their grouping as a multiword that allows the multiword elements to participate in the rest of the syntactic structure.

(3) [$_{\text{PN}}$ Jan Johansson] började spela piano 1942.
Jan Johansson started play.INF piano 1942
'Jan Johansson began to play the piano in 1942.'

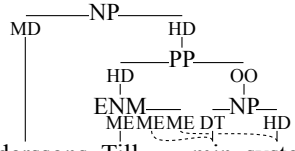
For cases like these, too, we need information about the presence of multiwords and their types as part of the syntactic structure. Without it, we would not be able to assemble the syntactic trees at all.

Eukalyptus therefore integrates multiword annotation into the syntactic annotation, using the possibilities of having secondary edges to be able to 'overlay' multiword annotation on top of regular syntactic structures. We recognize two types of multiwords: *Analyzed* multiwords are treated like just indicated: they receive a regular syntactic annotation, and in addition we insert a node with a multiword part-of-speech directly above one of the multiword parts in the primary graph, and link the other multiword parts to these nodes using secondary edges. *Unanalyzed* multiwords, on the other hand, are not considered to have syntactically meaningful internal structure, and their parts are therefore gathered under a multiword node in the primary graph. In both cases, the multiword node serves as the anchor of the SALDO sense id.

The examples in (1) and (2) above contain analyzed multiwords, their trees are given in (4) and (5) below. The special dependency label ME (multiword element) is used for the children of a multiword node. Note that the analyzed multiwords receive a regular syntactic analysis in the primary graph. The additional multiword verb node (VBM) above *dra* in the primary graph (4) can be considered to be superfluous from a syntactic point of view, all it does is provide an anchor for the SALDO id and connect the multiword elements. Since the TIGER/NEGRA formalism does not allow nodes that are only connected with secondary edges, this node has to appear somewhere in the primary graph. But although the multiword proper name node (ENM) above *till* in (5) is without effect in the primary graph directly surrounding it – for instance we still consider the preposition *till* to be the PP's head – it is instrumental when we check for violations of the headedness rules. In this case, we allow the PP to act as the head of an NP, since it's yield is also completely under an ENM node (in the full graph).

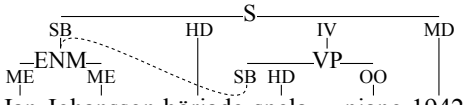


(4) Den andra slutade dra de allra tyngsta timmerstockarna .
the other stopped pull.INF the very heaviest timber logs.DEF
'The other one doesn't snore as heavily as he did before.'



(5) Anderssons Till min syster
Andersson's to my sister
'Andersson's (poem) *For my sister*'

The multiword proper name in (3) is an example of an unanalyzed multiword, its tree is given in (6). Note that in contrast to the previous two examples, the parts of an unanalyzed multiword are children of the multiword node in the primary graph, and the multiword elements are only marked with ME-function.



(6) Jan Johansson började spela piano 1942 .
Jan Johansson started play.INF piano 1942
'Jan Johansson began to play the piano in 1942.'

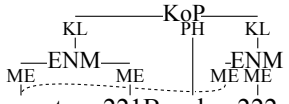
The analyzed-unanalyzed distinction is a type level rather than a token level distinction. As the status of being unanalyzed precludes any modification, and judging modifiability is, in our experience, unreliable, we try to treat as many multiwords as possible as analyzed. Of course, a central property of our scheme is that the choice for syntactical analysis is not mutually exclusive with recognition of its multiword status.

As unanalyzed multiwords we have for example discontinuous coordinators (*både . . . och* 'both . . . and'), circumpositions (*för . . . sedan* 'ago', lit. 'for . . . since'), compound numerals (*sju tusen femhundra* '7500'), phrases of foreign origin (*ad hoc*), and most person names (*Jan Johansson*) and addresses (*Bagaregatan 221B*).

As analyzed multiwords, we may mention adjective-noun combinations (*god man* 'agent with power of attorney', lit.: 'good man'), particle verbs (*gå bort* 'die', lit.: 'go away'), verb-argument idioms (*dra en vals* 'lie', lit.: 'turn a walz'; *gå på gatan* 'prostitute oneself', lit.: 'walk in the street'; *måla fan på väggen* 'assume the worst', lit.: 'paint the devil on the wall'), idiomatic coordinations (*vara ute och cykla* 'be confused/wrong', lit.:'be out and riding a bike'), proverbs (*Äpplet faller inte*

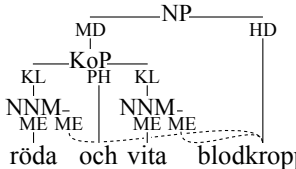
långt ifrån trädet ‘the apple doesn’t fall far from the tree’), analyzable proper names of different kinds (*Det sjunde inseglet* ‘The seventh seal’, *före detta jugoslaviska republiken Makedonien* ‘(the) former Yugoslav republic (of) Macedonia’), fixed PPs (*före detta* ‘former/ex-’, lit.:‘before this’), NP-formed date expressions (*den fjärde maj* ‘the fourth (of) May’), complex prepositions (*på grund av* ‘because of’, lit.: ‘on ground of’), and many more.

Together with the other Eukalyptus annotation principles, our treatment of multiwords is flexible enough to handle a great range of multiword types and uses, including elided multiword parts in coordinations – noted as a problem in [3]. Example (7) shows a coordination of two street addresses, with an elided streetname in the second conjunct. Street addresses are considered unanalyzed multiword proper names (ENM). In coordinations, we may thus see unanalyzed multiwords nodes that dominate some of their elements in the secondary, rather than the primary, graph. Note however, that nowhere in the graph do these elements enter the graph in a non-ME function, which means their inclusion in the graph is only licensed by virtue of their being multiword elements, which is the hallmark of an element in an unanalyzed multiword.



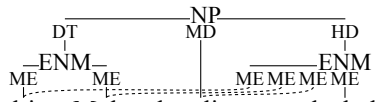
(7) Bagaregatan 221B och 222
Baker street 221B and 222
‘Baker street 221B and 222’

The example in (8) shows a coordinated multiword noun (NNM), analyzed as a coordination of adjectival attributes in an NP.



(8) röda och vita blodkroppar
red and white blood cells
‘red and white blood cells’

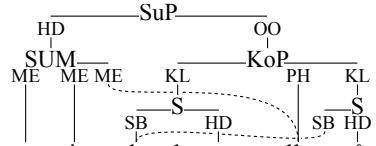
Furthermore, a strength of our approach is that analyzed multiwords can contain other multiwords, thus enabling us to handle embedding of multiwords such as proper names in titles:



(9) Adrian Moles hemliga dagbok
Adrian Mole's secret diary
‘The secret diary of Adrian Mole’

Thus far, we have come across one multiword that requires a split analysis, that is, it partially falls into the analyzed class and partially into the unanalyzed class. It concerns the multiword complementizer *vare sig* ... *eller* 'irrespective of whether ... or' (lit. 'be.SUBJ REFL ... or'). As shown in (10), the first two words together sit in complementizer position (head of subordinator phrase SuP), whilst the last word functions as coordinating conjunction inside the subordinate clause (pseudo-head PH of coordinator phrase KoP).

(10)



vare sig du kommer eller går
be REFL you come or go
'irrespective of whether you are coming or going'

A particular problem that shows up in our treatment of multiwords as syntactic units, and our decision to analyze multiwords and their parts as much as possible, is that multiword elements that do not have an independent usage may require ad hoc analyses. Take, for example, the multiword elements *slint* and *vika* of the idiomatic combinations *slå slint* 'fail, misfire' (lit. 'hit *slint*') and *ge vika* 'give way, give in' (lit. 'give *vika*') are not used in other contexts – even though we can easily trace their respective etymologies to the verbs *slinta* 'to slip' and *vika* 'to bend/yield/move'. We have chosen to treat these elements as nouns, because of the existance of other noun-verb pairs in Swedish whose forms relate to each other in the same way, and to treat the complete multiwords as verb-objekt idioms. But since these nouns never occur anywhere else than as (stipulated) objects to these verbs and they do not show object properties like fronting or promotion to subject in a passive, the analysis is not really meaningful. Treating multiwords as tokens, and thus as leaves in the syntactic tree would have avoided this forced classification. However, this would give rise to discontinuous tokens, which may be difficult to handle, visualize and reason about, and, more importantly, it would in essence reduce all multiwords to unanalyzed multiwords. We therefore feel the occasional need for ad hoc analysis is a fair price to pay.

The literature on multiwords, both in theoretical and computational linguistics, consists to a large part in setting up ontologies of multiwords, modelling the syntactic properties of different types of multiwords and investigating consequences for the formal grammar system. Seen against that background, our simple dichotomy may seem to be inadequate as it is nonrestrictive and does not necessarily provide any further insight into the nature of multiwords. However, as part of an annotation scheme, this is not only acceptable but arguably preferable. The task of a syntactic annotation scheme is to allow us to assign the structural distinctions of interest to a broad range of data, rather than to model the language in a generative sense. This is exactly what the analyzed-unanalyzed distinction allows us to do.

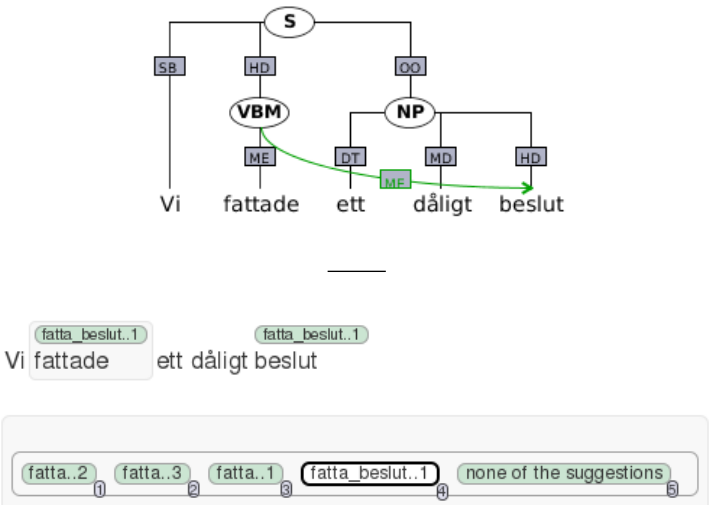


Figure 2: Annotating *vi fattade ett dåligt beslut* 'we made a bad decision' in the syntactic task (above) and in the word sense annotation task (below).

4 Multiwords in the Annotation Tasks

As mentioned above, multiwords occur in the word sense annotation as well as in the syntactic annotation. However, sense annotation and syntactic annotation require different annotation tools and methodologies, so for practical reasons we annotate these layers separately. The syntax annotators use a traditional treebank annotation tool,[1] and while their annotation guidelines describe how to treat multiwords, this tool is not integrated with the SALDO lexicon and does not help the annotators decide whether or not a multiword is present in the text. The sense annotators, on the other hand, use a sense annotation tool that is tightly integrated with SALDO, so that for each token, the annotator can choose from a list of single-word and multiword senses defined in SALDO. This makes it easier to know whether the lexicon defines a suitable multiword. We recognize that the subtask of detecting the presence of a multiword is essentially performed in both annotation tasks; however, these annotations will be harmonized in the final stages of the project. It also gives us the opportunity to investigate the influence of our tools and methodologies on this subtask.

Figure 2 shows an example of how a sentence is annotated using the syntactic and word sense annotation tools. In this sentence, *Vi fattade ett dåligt beslut* 'We made a bad decision', there is a discontinuous multiword *fatta . . . beslut* 'make . . . decision', which is annotated on the syntactic level using a node representing the

[1]The syntax tool is based on Synpathy, once developed but no longer maintained at the Max Planck Institute for Psycholinguistics, Nijmegen. See `http://spraakbanken.gu.se/koala` for more information.

multiword verb (VBM). In the word sense annotation tool, the annotator has to pick the multiword sense *fatta beslut*, rather than one of the senses of the single word entry *fatta* 'grasp; comprehend'.

We compared the multiword annotation in the parts of the treebank where syntactic and sense annotation were both complete; at the time of writing, this part consisted of 7,043 tokens. We did not evaluate how well the annotators were able to make the analyzable/unanalyzable distinction, since this distinction is made on the syntactic level only, nor did we evaluate the actual sense id selected, as this is only part of the sense annotation. The syntactic annotation layer contained 257 multiwords (excluding proper names) in this part of the corpus, while the sense layer had 374 multiwords. In 234 of these cases, the annotation was consistent between the layers, so the syntactic annotations had a precision of 0.91 and a recall of 0.63 with respect to the sense layer. This shows that there are few annotation conflicts: the syntactic annotation is more conservative, which is no doubt caused by the lack of lexicon integration in the syntactic annotation tool, and perhaps also by the required effort of inserting an extra multiword node in the syntactic tree in the case of analyzed multiwords. It is encouraging to see that *when* the syntactic annotators have a strong intuition that a multiword is present, it is also very likely to be annotated as a multiword on the sense level.

We finally considered the multiwords annotated in the sense layer but which were left out in the syntactic layer. As can be expected, they tend to belong to the category of analyzed multiwords, which are often harder to spot and which play a less central role in syntactic annotation. In particular, light verb constructions were often left out by the syntactic annotators (e.g. *fatta beslut* 'make decision' or *spela roll* 'play role'); these are among the syntactically most flexible, and thus inconspicous, of the multiwords.

5 Conclusions

We have shown how the Eukalyptus treebank of written Swedish handles the dual lexical and syntactic nature of multiwords, by formally locating them at the level of syntactic structure. We distinguish between two types of multiwords: analyzed multiwords, whose parts also have a regular syntactic role in the tree; and unalyzed ones, whose parts are only integrated by virtue of being in the multiword.

We are able to compare multiword detection by our lexical and our syntactic annotators. We see that the annotators agree well, however, it is clear that, in terms of tool support, integration of the lexical resource into the syntactic annotation work flow might improve detection of multiwords at that level. Note that, since it is straightforward to mechanically transfer the multiwords found during lexical annotation to the syntactic layer as analyzed multiwords, the lower recall of the syntactic annotators with respect to the lexical annotators is unproblematic. However, an issue for future investigation is how we may improve identification of multiwords that are not currently in the lexicon and are thus likely to be missed in both tasks.

Acknowledgements

The Koala project is funded 2014–2016 by Riksbankens Jubileumsfond, grant number In13-0320:1.

References

[1] Yvonne Adesam, Lars Borin, Gerlof Bouma, Markus Forsberg, and Richard Johansson. Koala – Korp's linguistic annotations. Developing an infrastructure for text-based research with high-quality annotations. In *Proceedings of the Fifth Swedish Language Technology Conference (SLTC)*, Uppsala, november 2014.

[2] Yvonne Adesam, Gerlof Bouma, and Richard Johansson. Defining the Eukalyptus forest – the Koala treebank of Swedish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 1–9, Vilnius, Lithuania, May 2015. Linköping University Electronic Press, Sweden.

[3] Eduard Bejček, Pavel Straňák, and Daniel Zeman. Influence of treebank design on representation of multiword expressions. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6608 of *Lecture Notes in Computer Science*, pages 1–14. Springer, Berlin, Heidelberg, 2011.

[4] Lars Borin, Markus Forsberg, and Lennart Lönngren. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211, 2013.

[5] Thorsten Brants, Roland Hendriks, Sabine Kramp, Brigitte Krenn, Cordula Preis, Wojciech Skut, and Hans Uszkoreit. Das NEGRA-Annotationsschema. Technical report, Universität des Saarlandes University, Dept of Computerlinguistik, Saarbrücken, 1999.

[6] Ulf Teleman. *Manual för grammatisk beskrivning av talad och skriven svenska*. Studentlitteratur, Lund, 1974.

[7] Ulf Teleman, Staffan Hellberg, and Erik Andersson. *Svenska Akademiens Grammatik*. Svenska Akademien, Stockholm, 1999.