Multiword Annotation in the Eukalyptus Treebank of Written Swedish

Gerlof Bouma Yvonne Adesam
Språkbanken / Dept of Swedish, University of Gothenburg
WG 4

1. About the Eukalyptus treebank

The *Eukalyptus treebank of written Swedish* is currently in an advanced stage of development, and will consist of ∼100k tokens of diverse text types such as blogs, Wikipedia, news, and novels. The treebank is used to evaluate automatic annotation tools for contemporary Swedish text as part of the Koala-project [1, 2].

Eukalyptus' syntax annotation scheme uses the NEGRA/TIGER format [3], combining (possibly discontinuous) phrases with labelled edges for the syntactic functions. Secondary edges are used to express sharing. In addition, we annotate morphological features, and lemmata and word senses from the lexical resource SALDO [4]. Segmentation/tokenization is based on orthographic word and sentence delimiters.

Multiwords lie at the intersection of the lexical and syntactic annotation levels. In addition, they are treated as first-class citizens in SALDO. Together, these facts give multiwords a prominent role in the multi-level annotation efforts in Koala.

2. Multiword annotation

At the heart of our multiword annotation scheme lies the decision to treat multiwords as part of syntax and not in a separate annotation layer. One of two annotation styles is used: First, multiwords that have internal structure analyzable in terms of the general Eukalyptus scheme are primarily treated as regular syntactic structures. An *additional* node is added to the graph representing the multiword, dominating its elements via secondary edges. These are the *analyzed multiwords*. Their trademark is that their multiword elements also play non-multiword-related roles in the syntactic tree. Secondly, multiwords that do not follow the general syntactic patterns receive a flat annotation. We gather the elements directly under a multiword node, which participates as a whole in the rest of the graph. We call these *unanalyzed multiwords*.

Incorporating multiword annotation in syntax is motivated by the desire to have a *single representation* that lets us a) annotate multiwords in addition to regular syntactic annotation, b) handle syntactically idiosyncratic structures, and c) account for multiword-related exceptions in our well-formedness criteria on headed structures. A positive consequence of our choice is that (re)use of a single token in several multiwords (e.g., coordination of multiwords with ellipsis, or embedding of one multiword within another) is not problematic [5].

Multiword nodes are essential in linking the treebank to the SALDO lexical resource. In SALDO, multiword entries are treated on a par with single word entries, and receive parts-of-speech and sense-ids. Our multiword node label matches the part-of-speech in SALDO, which is not determined by the internal structure of the syntactic object, but by the whole multiword's inflectional and distributional features. Analyzed multiwords may thus have two unrelated syntactic categories: one for the (regular) phrase and one for the multiword node. Idiomatic

PPs are common in this group of category shifters: these may be multiword adverbs (*med råge* 'abundantly', lit. 'with heap'), but also multiword proper names (titles: *(Malots) Utan släkt* '(Malot's) Sans famille'), interjections (*för fan!* 'dammit!', lit. 'for devil!'), adjectives (*på smällen* 'pregnant', lit. 'on the hit') or complementizers (*i den mån*, lit. 'to the extent').

The decision to treat a particular multiword as analyzed or unanalyzed is taken at type level. It is purely based upon what we consider to be the boundaries of the Eukalyptus scheme. For instance, we avoid positing ad hoc analyses just to shoe horn a type of expression into the regular scheme. So, person names (*Olof Palme*) and street addresses (*Bagaregatan 221B*) are unanalyzed: they follow a clear template or 'syntax' different from the general NP syntax. Compound numerals (*sju tusen femhundra* lit. 'seven thousand five-hundred') are unanalyzed for the same reason. Similarly, we see no principled basis for assigning a head-dependent relation to the parts of a circumposition (*for* ... *sedan* 'ago', lit. 'for ... since') or a discontinuous coordinator (*antingen* ... *eller* 'either ... or') – they are thus unanalyzed multiwords.

The dichotomy analyzed-unanalyzed does not directly correspond to any of the well-known multiword properties from the literature. Since Eukalyptus allows discontinuous nodes, the ability to move elements of the multiword around is not an argument for annotation as an analyzed multiword, cf. the discontinuous coordinators above. Conversely, being completely fixed is not an argument for an unanalyzed treatment – if the combination fits regular syntax, its syntactic structure will be annotated. A multiword preposition of the P-N-P template that is completely inflexible and doesn't allow internal modification, is still an analyzed multiword, as the PPs headed by them follow a normal syntactic pattern: $[_{\mathrm{PP}}\ \mathrm{P}\ [_{\mathrm{NP}}\ \mathrm{N}\ [_{\mathrm{PP}}\ \mathrm{P}\ [\ldots]\ ]\ ]\ ]$. Of course, internal modification is only possible in analyzed multiwords, as the flat annotation of unanalyzed multiwords lacks the structure needed to incorporate any modifiers in a meaningful way. In general, we have found judgements of internal modifiability to be unreliable. We therefore do not involve this criterion in the annotation, but rather treat as many multiwords as possible as analyzed.

3. Poster overview

In this poster we present the Eukalyptus treebank and its multiword annotation scheme to the PARSEME network. A recent survey of multiword annotation in 17 European language treebanks employs a typology of salient multiword types [6]. We will also use this typology to present different multiword types annotated in Eukalyptus. Because our multiword inventory starts from SALDO, which has a comprehensive view of multiwords, we can give examples from most of the types highlighted in said paper, and in addition many types not specifically mentioned there (i.e., from the 'other' category). The overview intends to make more clear our view of the connection between the dichotomy in our annotation scheme and the categorizations and concepts familiar from the literature, and to open for discussion of these connections.

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

[1] Y. Adesam, L. Borin, G. Bouma, M. Forsberg, and R. Johansson. Koala – Korp's linguistic annotations. Developing an infrastructure for text-based research with high-quality annotations. In *Proc SLTC*, Uppsala, 2014. [2] Y. Adesam, G. Bouma, and R. Johansson. Defining the Eukalyptus forest – the Koala treebank of Swedish. In *Proc NODALIDA*, pp 1–9, Vilnius, 2015. [3] Th. Brants, R. Hendriks, S. Kramp, B. Krenn, C. Preis, W. Skut, and H. Uszkoreit. Das NEGRA-Annotationsschema. TR Dpt Computerlinguistik, Saarbrücken, 1999. [4] L. Borin, M. Forsberg, and L. Lönngren. SALDO: a touch of yin to WordNet's yang. *LRE*, 47(4):1191–1211, 2013. [5] Y. Adesam, G. Bouma, and R. Johansson. Multiwords, word senses and multiword senses in the Eukalyptus treebank of written Swedish. In *Proc TLT 2015*, pp 3–12, 2015. [6] V. Rosén, G. Losnegaard, K. De Smedt, E. Bejček, A. Savary, A. Przepiórkowski, and V. Mititelu. A survey of multiword expressions in treebanks. In *Proc TLT 2015*, pp 179–193, 2015.