Multiword annotation in the Eukalyptus treebank of written Swedish

:: Multiword problems
- Multiwords combine phrasal and lexical properties, which breaks expectations. The annotation needs to handle both sides.
- Some multiwords defy syntactic analysis, others need structure to host modifiers etc.
- Multiword senses need an anchor, the multiword parts need to be related to this anchor.

:: Analyzed and unanalyzed multiwords in syntax

O.: Does the multiword fit the general syntactic picture for NPs, VPs, etc?
A.: sure! (preferred) ...analyzed

In both analyzed and unanalyzed multiwords, the multiword node is nonterminal, receives a (multiword) part-of-speech and anchors the sense id from SALDO.

:: Examples

<table>
<thead>
<tr>
<th>Analyzed</th>
<th>Unanalyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malot's</td>
<td>Malot's with out</td>
</tr>
<tr>
<td>Comp 2nd</td>
<td>&quot;Svensk&quot;</td>
</tr>
<tr>
<td>language</td>
<td>Swedish as a second</td>
</tr>
<tr>
<td>language</td>
<td>&quot;Swedish as a second</td>
</tr>
<tr>
<td>language</td>
<td>language</td>
</tr>
<tr>
<td>&quot;Gustav Vasa&quot;</td>
<td></td>
</tr>
<tr>
<td>&quot;Harry Potter&quot;</td>
<td></td>
</tr>
<tr>
<td>&quot;the Philosopher's Stone&quot;</td>
<td></td>
</tr>
<tr>
<td>1337 yrs</td>
<td>1337 yrs</td>
</tr>
</tbody>
</table>

:: Eukalyptus Treebank

Eukalyptus treebank, part of Koarp’s Linguistic Annotations (Koala) to improve the annotation toolchain underlying the corpus infrastructure at Spraakbanken.

Eukalyptus will contain ~100k tokens in five subcorpora: blog texts, fiction, wikipedia, Europarl, newspaper text/public information.

Annotations include lemmata, morphological properties, part-of-speech, syntactic structure, and word senses.

Because of the word sense annotation, multiwords have been a core part from the start.

:: TIGER style syntax

- Primary graph is tree formed.
- Phrases may be discontinuous.
- Edges carry function labels.
- Additional, secondary edges mark multiple roles.

:: SALDO word senses

Svensk assocationslexikon 2 (SALDO) is a Swedish lexical resource currently maintained and developed at Spraakbanken. It contains a full form lexicon, part-of-speech information and word senses defined by placement in an association network.

It contains about 8k multiword entries (~7%), treated on a par with single word entries.

:: Conclusions

Annotating multiwords as syntactic objects helps us to handle the Janus-faced character of many multiwords. We are able to tackle complicated cases like ellided multiword parts in coordination (not shown) and embedded multiwords (shown) for free.

We "double annotate", analyze, multiwords as much as possible, but some resist incorporation into the syntactic scheme. These are left unanalyzed. The resulting distinction should be seen in the context of the complete syntax or annotation scheme, and is not directly related to other well known multiword properties like compositionality and flexibility.