WG 4: ANNOTATING MWEs IN TREEBANKS

Summary from Athens Meeting

Working session 1
Monday March 10, 2014
16:30–18:00

Victoria Rosén began the session by reviewing the objectives for WG 4. The following were mentioned by Agata Savary in the introductory session of the meeting:

- guidelines for annotating MWEs in constituency and dependency treebanks
- extracting lexicons from treebanks
- extracting probability scores from treebanks

The Memorandum of Understanding lists two deliverables that are specific to treebanks and therefore central to the goals of WG 4:

- extensions of existing corpora and treebanks in several languages with MWE annotation levels
- annotation guidelines for the representation of MWEs in treebanks

The scientific work plan in the action proposal states that the main objective of WG 4 will be to take a step towards enhanced MWE-aware methodologies of treebank construction, and their optimal usability in parsing. It mentions the following as expected outcomes for WG 4:

- annotation guidelines for representing MWEs in constituency and dependency treebanks
- recommendations on how to use current and future treebanks to automatically extract lexicons and probability scores addressed in other WGs

Amongst the milestones for the action that are listed in the work plan, the following are especially relevant to WG 4:

- methodology of lexicon design accounting for MWEs and valence data
- methodology of MWE annotation in corpora
- meta-grammar formalism adapted to the representation of MWEs

The first year outcomes (from Agata's slides for the introductory sessions of the meeting):

- detailed scientific program of each WG
- contrastive state-of-the-art surveys in all WGs

Some of the above objectives can only be achieved by cooperating with other working groups. The objective which is most central to WG 4, at least in the early phase of the action, is the construction of guidelines for MWE annotation in treebanks. Such guidelines will probably need to be different for different types of treebanks and different annotation methods.

Some ideas for what such guidelines might address were mentioned.

- In a treebank you should be able see whether an expression is used compositionally or idiomatically. For example, the English phrase *a piece of cake* can be used as an idiom or literally. Which use is present in a sentence cannot be inferred in a simple way with statistical methods.
- MWEs that are 'words with spaces' should be marked as such, in other words it should be clear that they are inflexible. Example: *då och då* from the Swedish treebank is marked with the label *ID*, which apparently may be used for any idiom.
- Both contiguous and noncontiguous MWEs should be searchable, and their meaning as MWEs should be represented in some way.

There was agreement in the group that before we can start to address tasks like constructing annotation guidelines, we need to have an overview of the state of the art for MWE annotation in treebanks. The plan agreed on from the meeting in Warsaw was to look at examples of MWEs in various treebanks at the Athens meeting. Since a poster session was organized at this meeting, the poster presenters in WG 4 were asked to present the main ideas from their posters in the WG 4 session.

Gosse Bouma (the Netherlands): His goal is to see what is annotated in the Alpino Treebank. He observes a large diversity of annotations and types. He also reports on some arbitrary solutions. They treat as MWEs only patterns which are syntactically irregular, and there is no distinction between idioms and other MWEs with syntactic flexibility. A typology of MWEs is given.

Petya Osenova and Kiril Simov (Bulgaria): They present the current situation in BulTreeBank, which is a predominantly syntactic approach to MWEs. They also present the perspectives for the future. As such they see the combination of: selection-based, construction-based and catena-based approaches. The last approach operates on a sub-constituent level and is appropriate for idiosyncrasies.

Zdenka Uresova, Jana Sindlerova, Eva Fucikova and Jan Hajic (Czech Republic): They focus on verbal MWEs. In their approach the valency is central. The two layers (analytical and tectogrammatical) are connected in the MWE treatment. The work is set in a comparative context, since the translations of Czech MWEs into English are discussed.

Eduard Bejček, Pavel Straňák and Pavel Pecina (Czech Republic): They concentrate on annotation of MWEs in raw texts via connection to the parsed part of the same texts, and on the identification of the MWEs in parsed texts.

Veronika Vincze, István Nagy T. (Hungary): She focuses on the multilingual aspects of MWEs for German, English, Hungarian and French.

Victoria Rosén, Gyri Losnegaard and Koenraad De Smedt (Norway): The poster focuses on searchability within various treebanks. Consistent annotation is necessary in order for MWEs to be searchable. When there is no documentation, it is difficult to figure out how to search for possible MWE annotations, and the resulting guesswork takes a lot of time.

Working session 2
Tuesday March 11, 2014
9:00–10:30

The Tuesday session was opened with the last poster.

Dimitris Kokkinakis (Sweden): He presents the typology of MWEs. He needs them for events extraction. He works in the medical domain, where terminology prevails.

In the rest of the session we discussed how to proceed to produce an overview of MWE annotations in different treebanks. We decided that for each treebank in the overview, at least the following information should be included:

- theoretical framework
- availability, information about licensing
- version
- short description of how the treebank was (or is being) constructed
- static or dynamic
- URL link(s) to webpages, documentation, etc.

For each MWE type, the overview should include:

- an example sentence (or phrase, part of sentence), with glosses and an idiomatic translation
- an example representation from the treebank
- a prose description of the analysis, written in such a way that it can be understood by linguists working in different frameworks
- information about a search expression (and search tool) which can be used to find examples of the MWE in the treebank
- any problematic issues that need to be recorded

The types of MWEs to be included in the overview was discussed. We decided to use the types mentioned on Gyri's slides as a starting point.

- Verbs, nouns and adjectives with selected prepositions
- Particle verbs
- Verbs with both particle and selected preposition
- VP idioms
- Fixed expressions
- Named entities
- Complex numeral expressions

The following people volunteered to contribute to the overview:

- Eduard Bejček: Czech treebank
- Gosse Bouma: Dutch treebank
- António Branco: Portuguese treebank
- Matthieu Constant: French treebank
- Gyri Smørdal Losnegaard and Victoria Rosén: Norwegian, Swedish and English treebanks

- Kadri Muischnek: Estonian treebank
- Petya Osenova and Kiril Simov: BulTreeBank
- Veronika Vincze: English, Hungarian and German treebanks

Victoria and Gyri offered to find a suitable format for constructing and publishing the overview, and to make it available so that people can get started on providing the information before the next meeting. At the next meeting we will discuss whether changes need to be made before adding more MWE types and hopefully also more treebanks.