**purpose and aims**

Universal Dependencies, as a project to develop consistent grammatical annotation for dependency treebanks, create new opportunities for multilingual research and development in natural language processing, in areas like cross-linguistic evaluation of empirical results and multilingual parser development. However, since UD are not related to any particular language or language group, but indeed aim at creating a common annotation scheme for potentially any human language, it is quite possible that in a few cases some language specific information, encoded in local treebank annotation schemes, may be lost. In this project, the aim is therefore to answer the question if the use of UD has any impact on parsing quality in a monolingual environment, namely for Swedish. Since the main target of Universal Dependencies is in multilingual natural language processing, it may be worth investigating whether there are any costs or gains in using a UD annotated treebank in a situation where it would not be technically required.

**survey of the field**

In the field of natural language processing, and in that of syntactic parsing in particular, access to grammatically annotated treebanks is of key importance as of today. However, the annotation schemes of treebanks for different languages are often very different in structure – to the point where it is sometimes of considerable difficulty to say whether performance differences are to be explained by real structural divergence of languages or mere annotation differences between the treebanks (Nivre, 2015). Several steps towards a more consistent framework have been made in recent years.

In case of multilingual parsing, parallel corpora are frequently used. However, there have been some successful transfer attempts when parallel data is unavailable. These methods use resemblances in the dependency structure of languages. To encode these similarities, Neseem et al. (2010) uses a single set of universal rules for several different languages, guiding the grammar induction in the target language. As alternative to grammar induction in cross-lingual parser training, McDonald et al. (2011) shows a delexicalized direct transfer method, where for any training set only features like part-of-speech tags and syntactic attachment direction are used. The model is then built from the data of the annotated source language and is used to parse the target language. The authors note that the differences between annotation schemes in the treebanks are often the cause of the fact that some of the language pairs may work well together, while others – even if they are typologically similar – may sometimes not. Zeman et al. (2012) harmonize treebanks of 29 languages by means of mapping their annotation styles to a version of the scheme used by the Prague Dependency Treebank. Their method for automatic normalization of annotation styles applied transformation rules, conditioned on original annotation, dependency labels and morphosyntactic tags. Later, McDonald et al. (2013) show an improvement of the results of cross-lingual direct transfer parsing by using the Universal Treebank which contains a uniformed syntactic annotation scheme for several languages, thus enabling cross-lingual training of parser models. As a baseline for model transfer, delexicalized models are proposed. Experiments show, that even while parsers, trained on data from languages in the same language group, do achieve the best results, training parsers also across language groups is certainly not pointless. In order to achieve a more unified set of grammatical relations across languages, de Marneffe et al. (2014) propose a layered taxonomy of 42 relations, a list which can also be expanded by
language-specific ones. For native language identification, Swanson & Charniak (2014) develop data-driven methods for extraction of distinguishing linguistic features by comparison of Language 2 data to the Language 1 data of the writer, to identify features where L1 use also appears in L2, thus generating lists of candidate transfer features.

Recently, the project of Universal Dependencies has been gaining speed. It’s aim is to develop cross-lingual treebank annotation for a large number of languages. Being an extension of several previous efforts, its goal is to find unified approaches in regards to parts-of-speech, morphosyntactic descriptions and dependency relations (Zeman, 2015). To speed up the adoption, efforts are being made to convert the existing dependency treebanks to conform with Universal Dependencies. In case of Swedish, the widely used Swedish Treebank (Nivre & Megyesi, 2007) has been converted and is freely available in an updated version in the UD repository (Nivre, 2014).

**programme description**

The main goal of this project is to investigate if the use of Universal Dependencies has any impact on the parsing performance in comparison to the parsing results of the classic, non-UD, version of the Swedish Treebank (CSTB). In order to achieve this, test sets of the classic and the UD version (UDTB) of the Swedish Treebank will have to be parsed, with results evaluated and compared.

The project can be viewed as consisting of several chain links. After some initial theoretical preparation, parts of the parsing chain have to be set up. In order to acquire a somewhat more stable analysis picture, but also to minimize potential parser bias, the plan is to use 2 widely used dependency parsing systems, MaltParser (Nivre et al, 2007) and Stanford Parser (Klein & Manning, 2003).

For CSTB, an optimized parsing model is already available on the website of MaltParser. For Stanford Parser, such a model will have to be trained and optimized, something that can be achieved by testing various parameters on some 10 % of the training data, used as a development set. In case of UDTB, both MaltParser and Stanford Parser models will have to be created. For optimization of MaltParser’s UDTB model, MaltOptimizer tool can be used (Ballesteros & Nivre, 2012). For Stanford Parser, the UDTB procedure is similar to its CSTB case.

In order to actually parse both CSTB and UDTB sets, the test data will first have to be stripped from dependency relations and dependency relation tags, which can be done with a simple regular expression script. After parsing of both CSTB and UDTB test sets for each of the parsers, evaluation of results can commence.

Since representations in training sets of CSTB and UDTB are not equivalent, it is unreasonable to simply compare labelled attachment scores between the treebanks. Therefore, the current plan is to use the TedEval tool (Tsarfaty et al, 2011), whose evaluation metrics take into account different annotation schemes across multiple parsing experiments, providing a more objective measure of parsing performance. Finally, for each of the 2 parsers, CSTB and UDTB parsed data will be scored with the test data of each of the treebanks, giving some insights of how the use of Universal Dependencies impacts parsing of the Swedish Treebank.
Timeplan for the project is as follows:

<table>
<thead>
<tr>
<th>WEEK</th>
<th>TASK</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>theory/documentation study, parser environment setup</td>
</tr>
<tr>
<td>45</td>
<td>parser environment setup, data preparation</td>
</tr>
<tr>
<td>46</td>
<td>model training, parsing</td>
</tr>
<tr>
<td>47</td>
<td>parsing, evaluation, article</td>
</tr>
<tr>
<td>48</td>
<td>evaluation, article</td>
</tr>
<tr>
<td>49</td>
<td>article</td>
</tr>
<tr>
<td>50</td>
<td>article</td>
</tr>
</tbody>
</table>

**references**


