Towards a Uniform Representation of Treebanks: Providing Interoperability for Dependency Tree Data

Olga Pustylnikov
Department of
Computational Linguistics and
Texttechnology
Bielefeld University
D-33615 Bielefeld, Germany
Olga.Pustylnikov@uni-bielefeld.de

Alexander Mehler
Department of
Computational Linguistics and
Texttechnology
Bielefeld University
D-33615 Bielefeld, Germany
Alexander.Mehler@uni-bielefeld.de

Abstract

In this paper we present a corpus representation format which unifies the representation of a wide range of dependency treebanks within a single model. This approach provides interoperability and reusability of annotated syntactic data which in turn extends its applicability within various research contexts. We demonstrate our approach by means of dependency treebanks of 11 languages. Further, we perform a comparative quantitative analysis of these treebanks in order to demonstrate the interoperability of our approach.

1 Introduction

In recent years a large number of natural language resources providing structured information by means of annotated data have been developed. Among them, different types of corpora consisting, for example, of texts, multimodal documents, web documents or syntactic treebanks serve different scientific purposes and are available by means of specific schemata. To refer to these different types we speak of *corpus genres*.

Treebanks, instantiating a specific corpus genre, are syntactically annotated corpora which are mainly used in data oriented approaches to computational linguistics (Bod et al., 2003). The availability of these corpora is crucial for training and testing NLP applications as well as for exploring linguistic phenomena. Fortunately, a large number of syntactic treebanks is available for a multitude of languages.[1] However, these treebanks are provided in a wide range of different formats. That is, NLP tools as, e.g., syntactic parsers which are trained on a variety of languages in order to provide cross-lingual interoperability have to be adapted to ever new representation formats of such banks. Thus, a major problem of using treebanks in NLP is the high effort of adaptating the tools or transforming into the formats. A unification of existing formats reduces this effort and makes different treebanks applicable to divergent tools via a single interface. Although reusability of treebanks has a high priority (Kakkonen, 2005), mapping them onto a single format is not an easy task. This is explained with respect to three levels of corpus related features:

- **Level 1** refers to corpus genre related features.
- **Level 2** relates to specifics of the object data.
- **Level 3** includes features induced by the operative representation format.

On level 1 we distinguish, for example, between *dependency* and *constituency* structure-related treebanks. This distinction reflects different syntactic theories underlying the generation of the treebanks. Focusing on the corpus genre of dependency treebanks, they can be further distinguished with respect to the annotation requirements induced by the target language or by the specific dependency grammar in use.[2] This happens on level 2. On level 3 we distinguish formats of the target treebank (as, e.g., the

[1] See (Kakkonen, 2005) for a review on existing treebanks.

[2] See (Nivre, 2005) for a review on different dependency grammars.

Penn Treebank (Marcus et al., 1993), TUT (Bosco et al., 2000), NEGRA (Skut et al., 1998) or SUSANNE (Sampson, 1995))

In this paper we transform dependency treebanks of 11 languages into a single format in order to provide interoperability of cross-lingual NLP systems operating on them. For this task we take level 1, i.e. corpus genre-related, and level 3, i.e. format related differences into account. Thus, we present a format general enough to map dependency *and* constituency structures, but concentrate on dependency treebanks whose level 3 differences are eliminated. Note that we do not consider differences induced by the object data, that is level 2 features. The reason is that their elimination is more difficult (as in the case of dependency grammar-related differences) or even impossible (as in the case of language specific features). In summary, the present paper overcomes the deficit of a lacking representation format which maps the existing variety of treebanks and, thus, provides interoperability on the level of syntactic ontologies. We demonstrate this interoperability by a quantitative structure analysis, which – to the best of our knowledge – is the first one operating on 11 languages. Note that all freely available corpora being analyzed in this study can be downloaded from our web site.[3]

The paper is organized as follows: Its conceptual framework is described in Sec. 2. Sec. 3 presents our experimental setting which benefits from the unified representation of dependency treebanks. Sec. 4 presents the results of our comparative study. Finally, Sec. 5 discusses our findings while Sec. 6 gives a conclusion and prospects future work.

2 Towards a Unified Representation for Treebanks

Treebanks may differ with respect to genre, data, or format-related criteria as discussed in Sec. 1. It is highly desirable to reduce this variety in order to achieve better data access and reusability (Kakkonen, 2005). The starting point for providing portability of corpus data is to use XML as the primary format of data exchange. In the past, specific XML models were provided for representing instances of specific corpus genres, languages (and sometimes even of specific linguistic theories). An example is TIGER-XML which optimally fits syntactic treebanks but is not applicable to other types of corpora. In contrast to this, what we search for is a *generic* format allowing

- to integrate all kinds of treebanks and
- to be extensible to map the specifics of a given language or grammar.

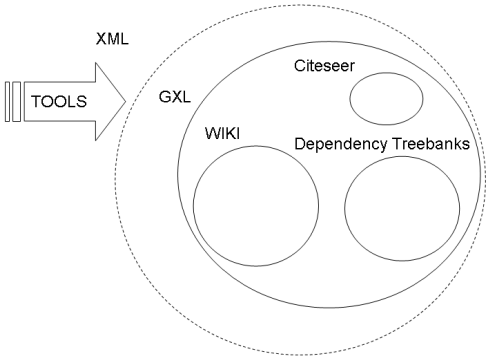

Figure 1: The Scope of GXL for corpus representation.

GXL (Holt et al., 2006) (see Figure 1) is an XML-based graph representation format which satisfies these two requirements as it allows one to deal with all kinds of graph structures.[4] Due to its generic graph data model it was successfully applied to various corpus genres such as Wikipedia-based corpora, newspaper corpora or dependency treebanks.[5] GXL represents corpus units as `node`-elements and their relations as `rel`- (or `edge`-) elements. All additional information of nodes (as, e.g., POS information) is stored within `attr`-elements. Each node is given an `id` so that it can be accessed via IDREF[6] attributes stored in `relend`-elements (in terms of the `target`- attribute – cf. Figure 2). The `direction` attributes (`in`/`out`) describe the kind of relationship between the nodes as, e.g., a head(`in`)-modifier(`out`) relation in a dependency tree.

[3] `http://ariadne.coli.uni-bielefeld.de/indogram/resources/`.

[4] See (Diestel, 2005) for the definition of a graph.

[5] See e.g. (Mehler and Gleim, 2005), (Mehler et al., 2007), (Ferrer i Cancho et al., 2007).

[6] IDentifier REFerence is a reference to a unique identifier in XML.

This general model allows the representation of various types of corpora and to store an arbitrarily large amount of additional information in order to account for their differences. That is, by mapping treebanks onto a GXL-based data model we enable tools to operate on different treebanks using the same interface and at the same time preserve their differences using a single representation format.

```
<rel>
  <relend direction="in" target="s30_7" />
  <relend direction="out" target="s30_8" />
</rel>
```

Figure 2: A dependency link between two nodes.

However, it turns out that this goal cannot be provided by GXL straightforwardly, but only by an extension of it henceforth called e*xtended* GXL (eGXL). The reason is that in many cases it is desirable to extend GXL to achieve a better fit to the object data. Due to its high level of generality, eGXL is less compact at some points. Consider, for example, the GXL-based representation of a token of the SUSANNE corpus (Sampson, 1995):

```
<node id="J02_p1_11">
    <attr name="lemma">
        <string>make</string>
    </attr>
    <attr name="pos">
        <string>VVNv</string>
    </attr>
    <attr name="word">
        <string>made</string>
    </attr>
</node>
```

Using GXL the token "made" is represented by 11 lines (3 lines per attribute). This is verbose from the point of view of data storage and retrieval. The specification of `attr` elements (`<string>...</string>`) is required from the schema allowing to store different data types like `string`, `integer`, etc. A more compact representation comparable to TIGER-XML encodes all extra information of a node by means of attributes in a single line. Thus, it should be possible to encode tokens from the SUSANNE corpus as follows:

```
<node id=".." form=".." lemma=".." />
```

One solution to achieve this is to extend GXL by means of additional node attributes providing a more compact storage of the data. Unfortunately, the original GXL-Schema[7] contains syntax errors making it impossible to derive from it. Thus, in a first step we corrected the errors[8] and extended the model by two `node` attributes, namely `form` and `lemma`[9]. Further, we added an attribute (`type`) to the `relend`-element to specify the type of relation. It allows one to define different types of relations once in the head of the document and access them by means of reference attributes.

What we get is a new graph representation scheme extendedGXL (eGXL)[10] which integrates the possibilities of TIGER-XML to represent syntactic trees. eGXL is extensible and can be adapted to more specific data while it remains generic being applicable to any kind of corpora.

```
<node id="Types">
  <graph id="g0">
    <node id="POS"/>
    <node id="t1" name="verb"/>
    <node id="t3" name="prepozitie"/>
    <node id="t5" name="substantiv"/>
    ...
    <node id="CAT"/>
    <node id="t2" name="subiect"/>
    <node id="t4" name="atribut subst."/>
    ...
    <edge from="POS" to="t1"/>
    <edge from="CAT" to="t2"/>
    <edge from="POS" to="t3"/>
    ...
  </graph>
</node>
```

Figure 3: eGXL *Types* graph.

```
<node id="Sentences">
   <graph id="g1">
     <node id="s1_1" form="Autorizatia" pos="t1" cat="t2"/>
     <node id="s1_2" form="pentru" pos="t3" cat="t4"/>
        ...
     <rel>
       <relend direction="in" target="s1_7"/>
       <relend direction="out" target="s1_1"/>
     </rel>
     <rel>
       <relend direction="in" target="s1_1"/>
       <relend direction="out" target="s1_2"/>
     </rel>
        ...
   </graph>
```

Figure 4: eGXL *Sentences* graph.

[7] `http://www.gupro.de/GXL/xmlschema/gxl-1.0.xsd`.

[8] See (Pustylnikov, 2007b) for details on error removal.

[9] Which seem to map important pieces of information since we observed them in almost all treebanks.

[10] `http://ariadne.coli.uni-bielefeld.de/indogram/resources/XML\%20Schemata/eGXL-1.0.xsd`.

```
1   Cathy   Cathy   N   N   eigen|ev|neut   2   su  _   _
2   zag zie V   V   trans|ovt|1of2of3|ev    0   ROOT    _   _
3   hen hen Pron    Pron    per|3|mv|datofacc   2   obj1    _   _
4   wild    wild    Adj Adj attr|stell|onverv   5   mod _   _
5   zwaaien zwaai   N   N   soort|mv|neut   2   vc  _   _
6   .   .   Punc    Punc    punt    5   punct   _   _

<sentence id="8" user="" date="">
 <word id="1" form="Detta" postag="POOP" head="2" deprel="OO"/>
 <word id="2" form="vill" postag="WVPS" head="0" deprel="ROOT"/>
 <word id="3" form="jag" postag="POPPHH" head="2" deprel="SS"/>
 <word id="4" form="bestämt" postag="AJ" head="2" deprel="AA"/>
 <word id="5" form="bemöta" postag="VVIV" head="2" deprel="VG"/>
 <word id="6" form="." postag="IP" head="2" deprel="IP"/>
</sentence>

************** FRASE ALB-2 **************
1 Valona (VALONA NOUN PROPER F Â$CITY) [1.10;VERB-SUBJ]
1.10 t [] (ESSERE VERB MAIN IND PRES INTRANS 3 SING) [0;TOP-VERB]
2 in (IN_MANO_A PREP POLI LOCUTION) [1.10;VERB-PREDCOMPL+SUBJ]
3 mano (IN_MANO_A PREP POLI LOCUTION) [2;CONTIN+LOCUT]
4 ai (IN_MANO_A PREP POLI LOCUTION) [3;CONTIN+LOCUT]
4.1 ai (IL ART DEF M PL) [2;PREP-ARG]
5 dimostranti (DIMOSTRANTE NOUN COMMON ALLVAL PL) [4.1;DET+DEF-ARG]
6 . (#\. PUNCT) [1.10;END]
```

Figure 5: 3 Treebanks: Dutch, Swedish and Italian.

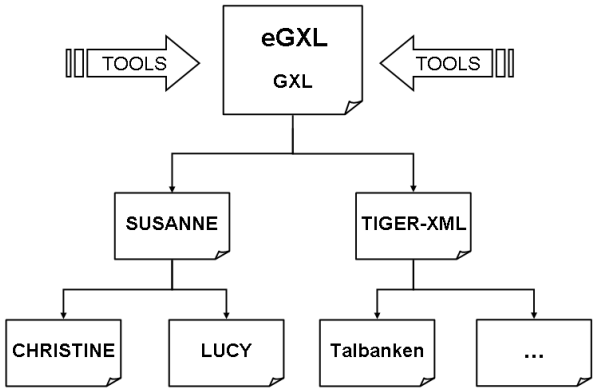

Figure 6: eGXL Hierarchy

2.1 The structure of eGXL

Figures 3 and 4 illustrate the general structure of eGXL. An eGXL document consists of two logical parts. The first part is the `Types` graph containing all attributes with the corresponding `ids`. That is, all possible values of an attribute (e.g. POS) are listed only once at the beginning of the document and accessed later by reference attributes. The second part consists of graph-based representations of sentences where the words are represented as `nodes` and their dependency relations as `rels`.

Generally speaking, treebanks vary with respect to content-related attributes (e.g., the `POS` attribute). Other than the `form` and `lemma` attributes, they are not part of the basic schema of eGXL. In order to map corpus specifics induced by such attributes we

```
<graph id="g1">
    <node id="s0_1" form="Cathy" lemma="Cathy" pos="t1" .../>
    <node id="s0_0"/>
    <node id="s0_2" form="zag" lemma="zie" pos="t4" extra="t4" ... />
    <node id="s0_3" form="hen" lemma="hen" pos="t7" extra="t7" ... />
    <node id="s0_4" form="wild" lemma="wild" pos="t10" .../>
    <node id="s0_5" form="zwaaien" lemma="zwaai" pos="t1" .../>
    <node id="s0_6" form="." lemma="." pos="t15" extra="t15" ...>
<rel>

<graph id="g8">
    <node id="s8_1" form="Detta" pos="t151" cat="t298"/>
    <node id="s8_2" form="vill" pos="t245" cat="t187"/>
    <node id="s8_0"/>
    <node id="s8_3" form="jag" pos="t152" cat="t306"/>
    <node id="s8_4" form="bestamt" pos="t26" cat="t254"/>
    <node id="s8_5" form="bemöta" pos="t227" cat="t312"/>
    <node id="s8_6" form="." pos="t86" cat="t86"/>
<rel>

<graph id="g2">
   <node id="n2_1" form="Valona" lemma="VALONA">
     <graph id="gn2_1">
       <edge from="n2_1" to="t16"/>
       <edge from="n2_1" to="t47"/>
       <edge from="n2_1" to="t7"/>
       <edge from="n2_1" to="t48"/>
     </graph>
   </node>
   <node id="n2_0" form="root"/>
   <node id="n2_1.10" form="t" lemma="ESSERE" />

<rel>
```

Figure 7: 3 Treebanks in eGXL: Dutch, Swedish and Italian.

provide a mechanism of extending eGXL which induces a family of XML Schemata all being derived from the basic eGXL schema (see Figure 6).

This model is not restricted to a particular treebank since all specific information of a treebank is instantiated by the `Types` graph. The core structure of the document remains always the same allowing the document to be accessed with the same tools. Figures 5 and 7 illustrate how three different input formats are transformed into a single representation. Figure 7 contains the sentences from Figure 5 transformed into eGXL. Although the sentences originate from different source treebanks they can be treated as a part of a single document regarding their structure.[11]

3 Quantitative Profiling of Dependency Treebanks

The unified representation format for dependency treebanks provided by eGXL allows us to compare 11 languages (Table 1) according to their quantitative characteristics. We benefit from our unified rep-

[11] Since the Italian treebank originally contains unlabeled attributes we include an additional attribute graph into a node element (Figure 7, the lowest part). This representation expands the node element, but the main shape of the document is preserved illustrating the extensibility of eGXL.

Treebank	Language	Size (#token)	Reference
Alpino Treebank v. 1.2	Dutch	195.069	(van der Beek et al., 2002)
Danish Dependency Treebank v. 1.0	Danish	100.008	(Kromann, 2003)
Sample of sentences of the Dependency Grammar Annotator	Romanian	36.150	`http://www.phobos.ro/roric/DGA/dga.html`
Russian National Corpus	Russian	253.734	(Boguslavsky et al., 2002)
A sample of the Slovene Dependency Treebank v. 0.4	Slovene	36.554	(Džeroski et al., 2006)
Talkbanken05 v. 1.1	Swedish	342.170	(Nivre et al., 2006)
Turin University Treebank v. 0.1	Italian	44.721	(Bosco et al., 2000)
CESS - Catalan Dependency Treebank	Catalan	100.000	(Civit et al., 2004)
Cast3LB - Spanish Dependency Treebank	Spanish	100.000	(Civit and Martí, 2005)
Prague Dependency Treebank 2.0	Czech	1.957.247	(Hajič, 1998)
BulTreeBank	Bulgarian	196.000	(Osenova and Simov, 2004)

Table 1: General Properties of the Treebanks.

resentation which provides a maximal reduction of level 3 differences (Sec. 1). The quantitative characteristics relate to dependency trees. The idea to compare languages by means of such features stems from (Ferrer i Cancho et al., 2004) who transformed 3 treebanks into Global Syntactic Dependency Networks (GSDNs) in order to measure their similarities. The nodes of GSDNs model are tokens where edges occur between two nodes if there is at least one dependency link between the corresponding tokens in the input bank. (Ferrer i Cancho et al., 2007) found out that the GSDNs of seven languages exhibit similar network properties which seem to be possibly universal properties of these kinds of networks.

Obviously, treebanks cannot be distinguished in terms of such measures. Thus, we focus on a different set of their structural characteristics. The aim is to answer the following questions:

- Can we classify treebanks by means of quantitative properties?
- Does the explored classification relate to known differences of the languages being analyzed?

In summary, we treat the above questions as a classification task in terms of *quantitative structure analysis* (Pustylnikov, 2007a; Mehler et al., 2007) using feature vectors to represent structural properties of treebanks.

3.1 Quantitative Dependency Tree Characteristics

We treat dependency trees of sentences as the basic unit to compute the characteristics listed below for each of the 11 input corpora. In order to get a single value of these characteristics for each of the corpora we average over all sentence-related observations of the respective corpus. The quantitative characteristics being computed are defined as follows:

In and Out Degree: The in (out) degree is given by the number of outgoing (incoming) dependency links observed for each word in the corpus.

Sentence Length: The sentence length is the average sentence length of a treebank.

Depth: The depth is the average depth of a dependency tree (sentence).

Depth Imbalance: As a measure of the imbalance of the sentence trees of a treebank we compute their A*bsolute* D*epth* I*mbalance* (ADI) according to (Botafogo et al., 1992). Starting from an input vertex v, this measure basically computes the standard deviation of the adjusted heights of v's child nodes. We compute the ADI for the root vertex r of the sentence tree T of each dependency treebank, where the higher $\mathrm{ADI}(r) \in \mathbb{R}_+$ the higher the variance among

the heights of r's child nodes, the more imbalanced T.

Child Imbalance: By analogy to the ADI we also compute the A*bsolute* C*hild* I*mbalance* (ACI) (Botafogo et al., 1992). Whereas the ADI evaluates imbalance in terms of the heights of child nodes, the ACI focuses on the sizes of the trees dominated by these nodes. Size is measured as the number of vertices of the respective tree. Obviously, the ADI also reflects the width of a tree and, thus, provides complementary information to the ACI.

Compactness and Stratum: Finally, the stratum and the compactness measures – as introduced by (Botafogo et al., 1992) – operate on graphs. In the present case we apply the measures to sentence trees which can be described as subsets of graphs. The Stra*tum* (Stra) is a metric which measures, so to speak, the deviation of a given sentence graph (tree) from a purely linearly organized graph with the same number of vertices where a stratum of 1 indicates a maximally hierarchically organized sentence. The C*ompactness* (C) analogously varies from 0 (i.e. graphs that are completely disconnected) to 1 (i.e. graphs that correspond to completely connected graphs). In our case the maximal values of 0 and 1 are never achieved since we deal with trees, which in turn are never completely connected or disconnected. Nevertheless, we expect the (C) values to vary for the different dependency treebanks reflecting different sentence structures.

3.2 Quantitative Structure Analysis

In text classification, structural features revealed to be a good alternative to the traditional *bag of words* approach (Pustylnikov and Mehler, 2007; Mehler et al., 2007). To build the feature vectors for our language-related classification task we compute the values of the characteristics listed in Sec. 3.1 for the 11 treebanks after being transformed into the eGXL. The aggregation of the feature values was done by computing the mean, the standard deviation and the entropy of the corresponding corpus-related value distributions. Each treebank is finally represented by a numerical vector with the cardinality $M \times N$ where M represents the number of characteristics and $N = 3$ is the number of location and dispersion parameters in use. To classify the treebanks we use semi-supervised hierarchical clustering. The results obtained in the experiment are presented in Sec. 4.

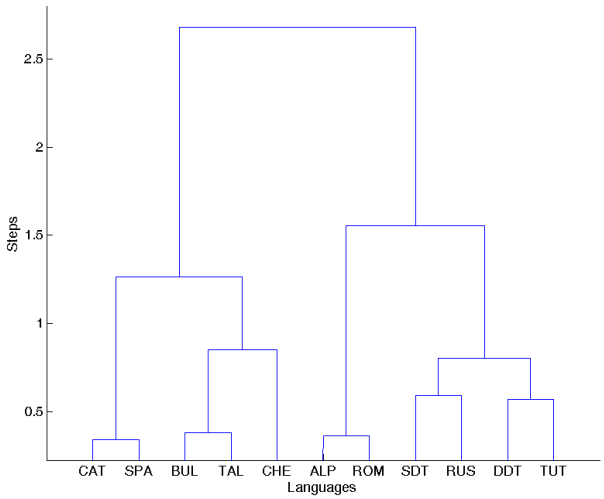

Figure 8: Results from language clustering. SDT - Slovenian, ROM - Romanian, RUS - Russian, DDT - Danish, ALP - Dutch, SPA - Spanish, TUT - Italian, BUL - Bulgarian, TAL - Swedish, CAT - Catalan, CHE - Czech.

4 Experimental Results

In Figure 4 the clustering results are visualized by a dendrogram. Each associated pair expresses the maximal similarity between two clusters among all clusters in every step. Thus, on the bottom the most similar languages according to the treebank characteristics are combined. The similarity threshold increases iteratively so that less similar clusters become connected. Finally, all groups constitute a single cluster (*bottom-up* clustering).[12] The height of the connection between two clusters indicates the strength of the similarity between them, that means, the higher the connection, the less similar are the two clusters to each other (Manning and Schütze, 1999).

In our case, Dutch (ALP) and Romanian (ROM) or Spanish (SPA) and Catalan (CAT) exhibit the

[12] Alternatively we can start from one cluster including all languages and divide them by reducing the similarity threshold in every iteration step (*top-down* clustering).

greatest similarity (the lowest connections) whereas the least similarity is expressed by means of the highest connection combining all languages within one cluster. A detailed discussion of the findings is presented in Section 5.

5 Discussion

At first glance, the overall partition of languages into clusters is far from their genetic classification (i.g. germanic, romance, slavic etc.). Looking at the similarity threshold of around 1.5 in the middle of the dendrogram we can point out three clusters differing internally in size and in similarity degrees. The first cluster contains SPA and CAT which (together with ROM and ALP) have the lowest connection as well as Bulgarian (BUL), Swedish (TAL) and Czech (CHE). The second cluster consists of two languages: ROM and ALP which are also dissimilar to other languages since they become merged with the third cluster only around the threshold of 1.6. The third cluster combines Slovenian (SDT) and Russian (RUS) and Danish (DDT) and Italian (TUT). The close connection between RUS and SDT is in accordance with our intuition about the membership of both languages in the slavic family. Similarly, SPA and CAT which are closely related genetically are also grouped together. Connections between languages which cannot be attributed to their genetic relationships require other explanations. Obviously, languages which differ genetically can nevertheless share structural properties (e.g. with respect to syntax or morphology). Since the observations we make about languages are related to dependency structures there are many possible reasons letting the sentence structure exhibit a particular shape. Italian for example is a Romance language which has preserved its inflectional morphology which may relate it to Slavic languages with respect to its structure. Bulgarian, a South-Slavic language has a tendency towards isolating / agglutinating languages which makes it group together with Swedish (TAL) and Czech (CHE). To complete the picture and to verify the assumptions further investigations need to be carried out which are beyond the scope of the present study. Here, we aimed at illustrating the possibilities for comparative investigations which arise with a unification of corpora.

6 Conclusion

In this paper we introduced a new XML based format for treebank representation. The format corrects and extends the generic graph model GXL to provide an effective means of integrating additional information in terms of `node` attributes. Our main goals have been

a) to provide a unification of 11 dependency treebanks,

b) to develop a representation format allowing to integrate the peculiarities of particular treebanks and

c) to combine the benefits of existing formats like TIGER-XML, etc. within a single representation.

We illustrated the potential of eGXL-based dependency treebank representations by a quantitative study. A unification of treebanks developed under different conditions is a demanding task with respect to format, language and annotation specific differences. A unification on the level of format by means of eGXL enabled the comparative quantitative investigation of 11 languages with a elevenfold reduction in computation effort. The results are in part interpretable in terms of language typology as in case of Slovene and Russian as well as of Spanish and Catalan. A systematic study of the impact of quantitative characteristics of dependency trees on language classification will be part of future work.

References

Rens Bod, Remko Scha, and Khalil Sima'an, editors. 2003. *Data-Oriented Parsing*. CSLI Publications, Stanford.

Igor Boguslavsky, Ivan Chardin, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Iomdin, Leonid Kreidlin, and Nadezhda Frid. 2002. Development of a dependency treebank for russian and its possible applications in NLP. In *Proc of LREC 2002*.

Cristina Bosco, Vincenzo Lombardo, Daniela Vassallo, and Lesmo Lesmo. 2000. Building a treebank for Italian: a data-driven annotation schema. In *Proc. of LREC 2000*.

Rodrigo A. Botafogo, E. Rivlin, and B. Shneiderman. 1992. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180.

Montserrat Civit and M.A. Martí. 2005. Building Cast3LB: A Spanish Treebank, a Research on Language and Computation. *Springer Verlag*, pages 549–574.

M. Civit, N. Bufí i, and P. Valverde. 2004. CAT3LB: a Treebank for Catalan with Word Sense Annotation. In *TLT2004*, pages 27–38. Tubingen University.

Reinhard Diestel. 2005. *Graph Theory*. Springer, Heidelberg.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtský, and Andreja Žele. 2006. Towards a Slovene dependency treebank. In *Proc. of LREC 2006*.

Ramon Ferrer i Cancho, Ricard V. Solé, and Reinhard Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review E*, 69:051915.

Ramon Ferrer i Cancho, Alexander Mehler, Olga Pustylnikov, and Albert Díaz-Guilera. 2007. Correlations in the organization of large-scale syntactic dependency networks. In *TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, pages 65–72.

Jan Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In E. Hajičová, editor, *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132. Karolinum, Charles University Press, Prague, Czech Republic.

Richard C. Holt, Andy Schürr, Susan Elliott Sim, and Andreas Winter. 2006. GXL: A graph-based standard exchange format for reengineering. *Science of Computer Programming*, 60(2):149–170.

Tuomo Kakkonen. 2005. Dependency Treebanks: Methods, Annotation Schemes and Tools. In *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, pages 94–104, Joensuu, Finland.

Matthias T. Kromann. 2003. The danish dependency treebank and the underlying linguistic theory. In Joakim Nivre and Erhard Hinrichs, editors, *Proc. of TLT 2003*. Växjö University Press.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. In *Computational Linguistics 19*.

Alexander Mehler and Rüdiger Gleim. 2005. The net for the graphs – towards webgenre representation for corpus linguistic studies. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as corpus*, pages 191–224. Gedit, Bologna, Italy.

Alexander Mehler, Peter Geibel, and Olga Pustylnikov. 2007. Structural Classifiers of Text Types: Towards a Novel Model of Text Representation. *To appear in: LDV Forum*.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proc. of LREC 2006*.

Joakim Nivre. 2005. Dependency Grammar and Dependency Parsing. Technical Report MSI report 05133, Växjö University: School of Mathematics and Systems Engineering.

Petya Osenova and Kiril Simov. 2004. BTB-TR05: BulTreeBank Stylebook. BulTreeBank Project Technical Report Nr. 05. Technical report, Linguistic Modelling Laboratory, Bulgarian Academy of Sciences.

Olga Pustylnikov and Alexander Mehler. 2007. Structural differentiae of text types. a quantitative model. In *Proceedings of the 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKl)*.

Olga Pustylnikov. 2007a. Guessing Text Type by Structure. In *Proceedings of the ESSLLI Student Session '07*, pages 221–231.

Olga Pustylnikov. 2007b. GXL-extension (GXXL-1.0). Correcting errors in GXL.

Geoffrey Sampson. 1995. *English for the Computer*. Clarendon Press, Oxford.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A Linguistically Interpreted Corpus of German Newspaper Text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany.

Leonoor van der Beek, Gosse Bouma, Robert Malouf, and Gertjan van Noord. 2002. The Alpino dependency treebank. In *Computational Linguistics in the Netherlands CLIN*, Radopi.