Building Parallel Treebanks for the Lesser-Resourced Languages.

A report for the Clarin-D center

Oleg Kapanadze

E-mail: ok@caucasus.net

Abstract

Naturally-occurring text in many languages are annotated for linguistic structure. A Treebank is a text corpus in which each sentence has been annotated with syntactic structure. Treebanks are often created on top of a corpus that has already been annotated with part-of-speech tags. The annotation can vary from constituent to dependency or tecto-grammatical structures. Treebanks have become valuable resources as repositories for linguistic research.

In this report we describe our experimental undertaking on building parallel Treebanks for German-Georgian, German-Russian and German-Ukrainian language pairs. The languages (except German) involved in the project from the computational viewpoint are considered "lesser-resourced" languages.

The parallel Treebanks can be used in translation studies, in corpus linguistics for studying syntactic phenomena, in computational linguistics as evaluation corpora for different NLT systems or for training and testing parsers and as a database for Translation Memory systems.

1. Introduction

Parallel corpora are language resources that contain texts and their translations, where the texts, paragraphs, sentences, and words are linked to each other. In the past decades they became useful not only for NLP applications, such as machine translation and multi-lingual lexicography, but are considered also very useful for empirical language research in contrastive and translation studies.

Naturally-occurring text in many languages are annotated for linguistic structure. A Treebank is a text corpus in which each sentence has been annotated with syntactic structure. Treebanks are often created on top of a corpus that has already been annotated with part-of-speech tags. The annotation can vary from constituent to dependency or tecto-grammatical structures. In turn, Treebanks are sometimes enhanced with semantic or other linguistic information and are skeletal parses of sentences showing rough syntactic and semantic information.

Treebanks have become valuable resources as repositories for linguistic research. They can be used in translation studies, in corpus linguistics for studying syntactic phenomena, in computational linguistics as evaluation corpora for different NLT systems or for training and testing parsers.

In this report a work on building parallel Treebanks for the language pairs German-Georgian, German-Russian and German-Ukrainian is outlined. From the computational view point three of the mentioned languages, except the German Language, are considered to be "the lesser-

resourced languages". Besides, typologically German and Georgian is much more dissimilar language pair, than the rest two pairs are.
The objective of the mentioned visit was not development of the full-scale parallel treebanks for the three languages pairs which would be unrealistic given the short notice of the research stay at the University of Saarland. Rather, the aim was starting with simple sentences in all four languages

- to tag and lemmatize manually terminal nodes
- produce syntactic parses for monolingual parallel German, Georgian, Russian and Ukrainian resources
- compare nonterminal nodes for determining stable and possible equivalents between phrases across the syntactic structures of the languages involved
- establish compatible tag-sets for Georgian, Russian and Ukrainian and , if necessary, to introduce the new syntactic phrase categories.

On the ground of the developed monolingual resources the further objective of the experiment envisioned

- production of the parallel trees for the bilingual resources
- alignment of the German-Georgian, German-Russian and German-Ukrainian parallel trees
- making general conclusions concerning feasibility of development treebanks for the mentioned language pairs.

2. Resources for experiment

For the low-density languages, including Georgian, Russian and Ukrainian, parallel corpora are very rare. The parallel texts used for the outlined experiment comprises German sentences and their translations into Georgian and Russian languages compiled for the GREG NLP lexicon project (Kapanadze et al., 2002, Kapanadze, 2010). The GREG itself contains valency data with the manually aligned Georgian, Russian, English and German verbs (ca. 1250) augmented with the examples of sentences considered as translation equivalents. Each subcorpus used for the study has a size of roughly 2600 sentence pairs that correspond to different syntactic subcategorization frames considered as German-Georgian translation equivalents. For the Russian and Ukrainian languages translation equivalents were provided by Dr. Alla Mishchenko, a DAAD postdoctoral fellow at the University of Saarland. She also took an active part in development of the monolingual Russian and Ukrainian resources and alignment procedures of the bilingual German-Russian and German-Ukrainian treebanks.

3. Building Monolingual Treebanks

3.1. Morphological analysis.

Initially emphasis has been made on development of a parallel treebank for a typologically dissimilar language pair German and Georgian, since the later is an agglutinative language using both suffixing and prefixing. For the Georgian text analyses has been applied a finite-state morphological transducer using the XEROX FST tools (Kapanadze 2010a,b), (Kapanadze 2009). The Georgian FST transducer utilizes a number of the formalisms supported by the XEROX toolkit (Beesley and Karttunen, 2003). The lexicon specification language lexc was used for modeling the lexicon and for constraining the morphotactics. It consists of 7 modules

for noun, adjective, pronoun, numeral, adverb, verb and the minor categories analysis. Currently there are two versions of the Georgian FST transducer available in the MS Windows platform and in the LINUX UBUNTU version.

For the rest of languages, German, Russian and Ukrainian, involved in the experiment, morphological features, including POS tags, were assigned manually drawing on the TIGER guidelines for the German language with the necessary changes relevant to the Russian and Ukrainian grammar formal description.

3.2 Syntactic parsing

The syntactic annotation employs parts-of-speech tags, morphological properties, and dependency functions. Every sentence is assumed to have a unique head and all other tokens, except punctuation marks, are direct or indirect dependents of the head. Monolingual files are XML-formatted.

Using the morphologically annotated bilingual corpus for each pair (German-Georgian, German-Russian and German-Ukrainian) the syntactical annotation were done manually. For this purpose we utilised the Synpathy, a tool for syntactical annotation developed at Max Plank Institute for Psycholinguistics, Nijmegen, the Netherlands (www.mpi.nl/corpus/manuals/manual-synpathy.pdf), a CLARIN-D project collaborator.

The German treebank annotation follows the TIGER annotation scheme (Skut et al., 1997, Brants et al., 2002). The other theree monolingual treebank were annotated according an adapted version of the German TIGER guidelines. The output of the syntactic annotation is in the TIGER-XML format. From the TIGER-XML format, the syntactic annotation may be visualized with tools like TIGER Search, representing dependency graphs for sentences. In Figure 1 and Figure 2 are examples of dependency trees for German, Georgian and Russian Sentences.

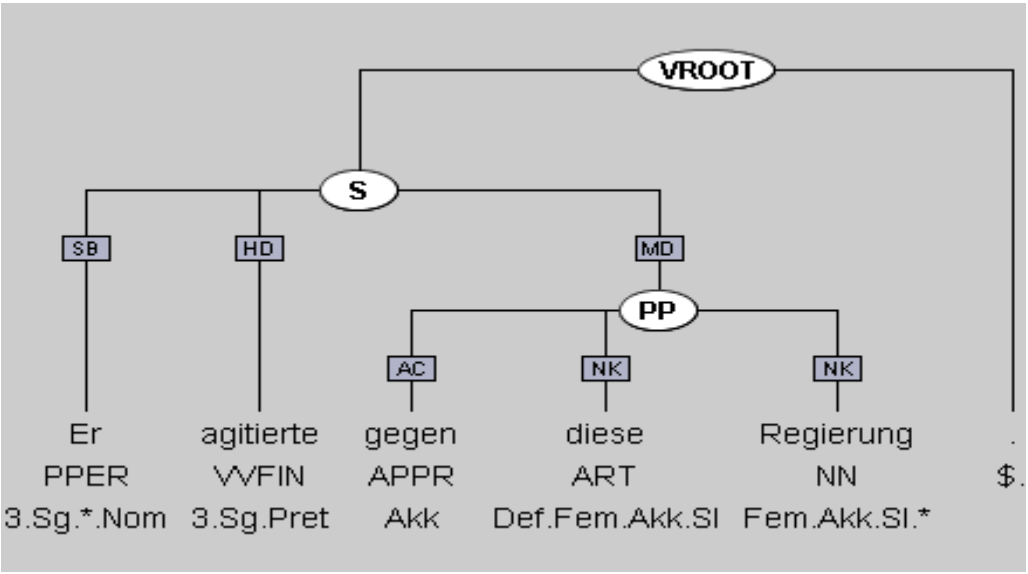

Figure1: A screenshot of an annotated sentence in German language.

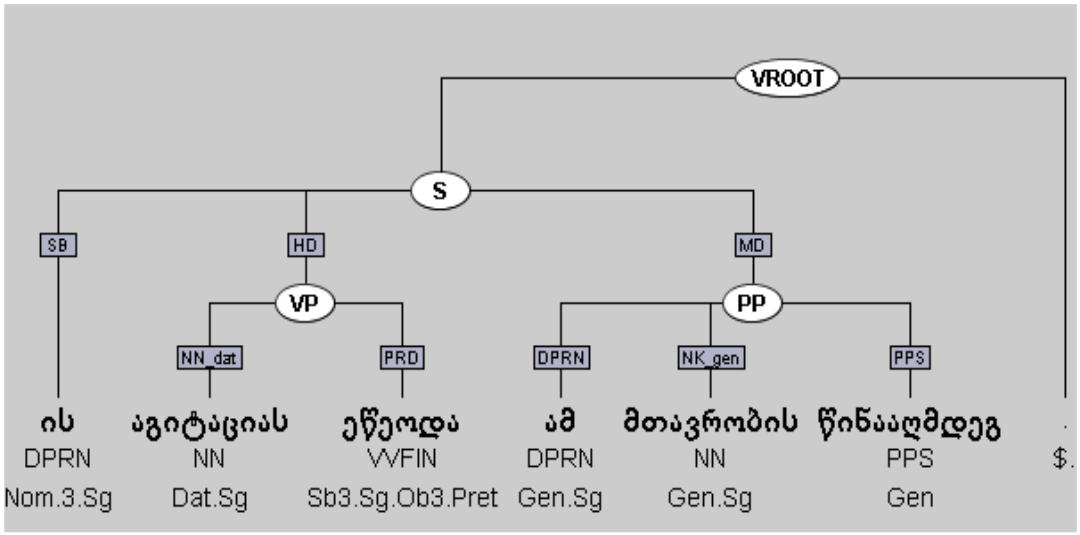

Figure 2: A screenshot of the corresponding annotated Georgian sentence.

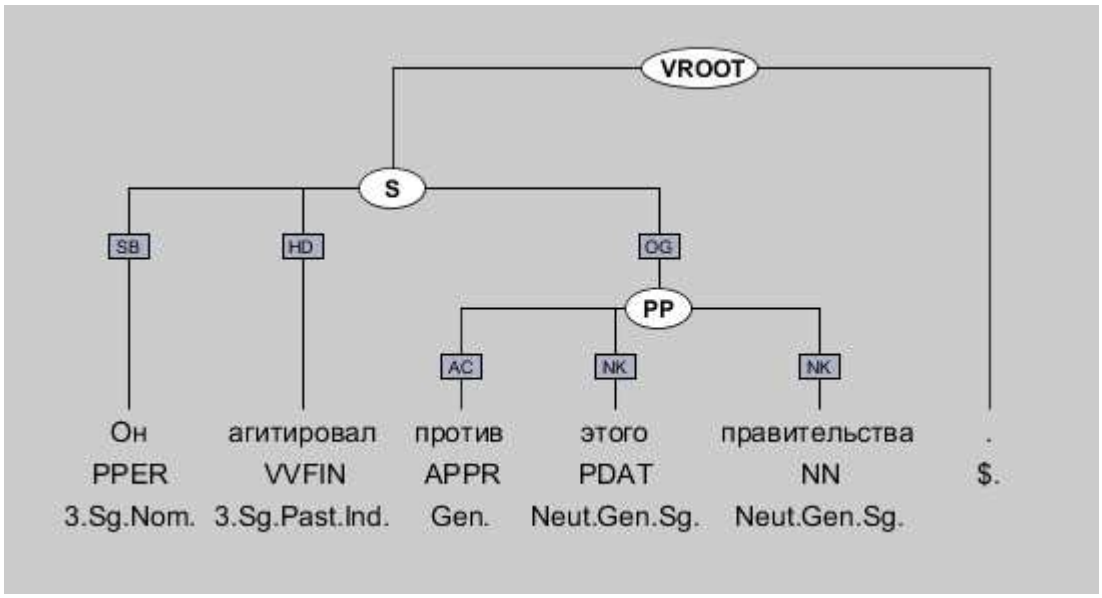

Figure 3: A screenshot of the corresponding annotated Russian sentence.

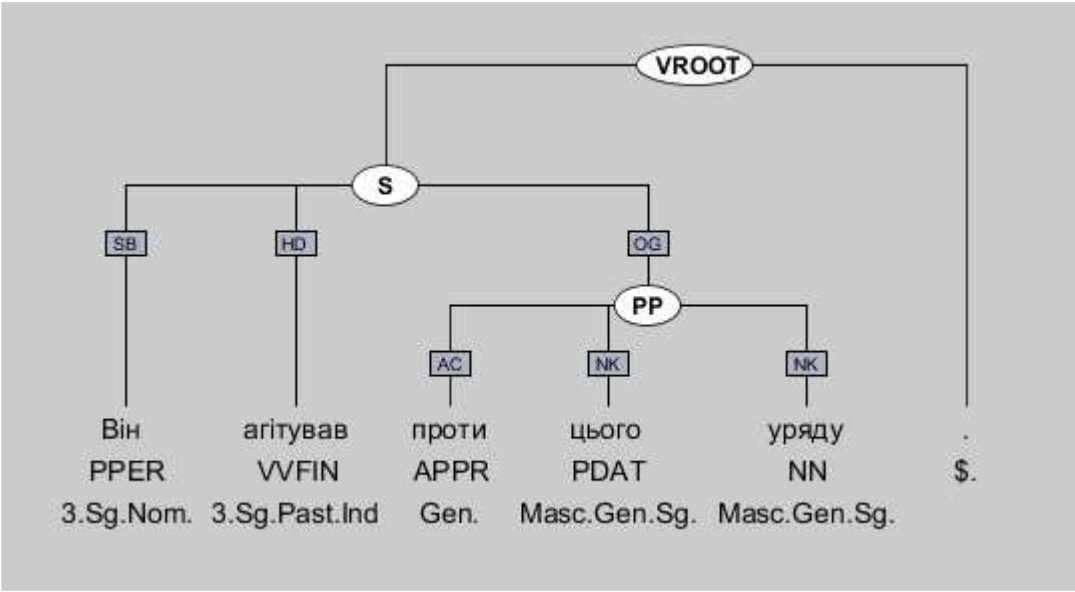

Figure 4: A screenshot of the corresponding annotated Ukrainian sentence.

The monolingual treebanks converted into TIGER-XML, are a powerful database-oriented representation for graph structures. In a TIGER-XML graph each leaf (= token) and each node (= linguistic constituent) has a unique identifier (Samuelsson and Volk, 2007). We use these unique identifiers for the phrase and word alignment across trees in corresponding translation units.

An XML representation is also used for storing this alignment. In the Figure 5 there is a representation of the Georgian sentence from the Figure 2 in the TIGER-XML format.

```
<body>
<s id="s12">
 <graph root="s12_502" discontinuous="true">

 <terminals>
    <t id="s12_1" word="ის" pos="DPRN"
     morph="Nom.3.Sg" />
    <t id="s12_2" word="აგიტაციას" pos="NN"
     morph="Dat.Sg" />
    <t id="s12_3" word="ეწეოდა" pos="VVFIN"
     morph="Sb3.Sg.Ob3.Pret" />
    <t id="s12_4" word="ამ" pos="DPRN"
     morph="Gen.Sg" />
    <t id="s12_5" word="მთავრობის" pos="NN"
     morph="Gen.Sg" />
    <t id="s12_6" word="წინააღმდეგ" pos="PPS"
     morph="Gen" />
     <t id="s12_7" word="." pos="$." morph="--" />
</terminals>

 <nonterminals>
```

```
  <nt id="s12_502" cat="S">
     <edge label="SB" idref="s12_1" />
     <edge label="HD" idref="s12_503" />
     <edge label="MD" idref="s12_501" />
  </nt>
  <nt id="s12_503" cat="VP">
     <edge label="NN_dat" idref="s12_2" />
     <edge label="PRD" idref="s12_3" />
  </nt>
  <nt id="s12_501" cat="PP">
     <edge label="DPRN" idref="s12_4" />
     <edge label="NK_gen" idref="s12_5" />
     <edge label="PPS" idref="s12_6" />
  </nt>
<nt id="s12_VROOT" cat="VROOT">
     <edge label="--" idref="s12_502" />
     <edge label="--" idref="s12_6" />
 </nt>
</nonterminals>
</graph>
</s>
```

Figure 5: A TIGER-XML format representation of a Georgian sentence from the Figure 1.

4. Building Parallel Treebanks.

Alignment of a Monolingual German, Georgian, Russian and Ukrainian Treebanks into a Parallel Treebank

This procedure is done with help of the Stockholm TreeAligner, a tool for work with parallel treebanks which inserts alignments between pairs of syntax trees (Samuelsson and Volk, 2005, Samuelsson and Volk, 2006). The Stockholm TreeAligner handles alignment of tree structures, in addition to word alignment, which – according to its developers - is unique (Samuelsson and Volk, 2006).

Phrase alignment can be regarded as an additional layer of information on top of the syntax structure. It shows which part of a sentence in the German language is equivalent to which part of a corresponding sentence in the other language. This is done with the help of a graphical user interface of the Stockholm TreeAligner. We drew alignment lines manually between pairs of sentences, phrases and words over parallel syntax trees. Figure 6 shows a screenshot with two aligned trees from Figure 1 and Figure 2. We intended to align as many phrases as possible. The goal is to show translation equivalence. Phrases shall be aligned only if the tokens, that they span, represent the same meaning and if they could serve as translation units outside the current sentence context. The grammatical forms of the phrases need not fit in other contexts, but the meaning has to fit.

The Stockholm TreeAligner guidelines allow phrase alignments within m : n sentence alignments and 1 : n phrase alignments. Even though m : n phrase alignments are technically possible, we have only used 1 : n phrase alignments, for simplicity and clarity reasons. One example of 1: n alignment on the word level is the Georgian multi-word expression for

“აგიტაციის გაწევა” represented under a VP node in the Figure 2, which is one word (“agitierte”) in the corresponding German sentence from the Figure 1. The 1 : n alignment option is not used if a node from one tree is realized twice in the corresponding tree, e.g. a repeated subject in coordinated sentences.

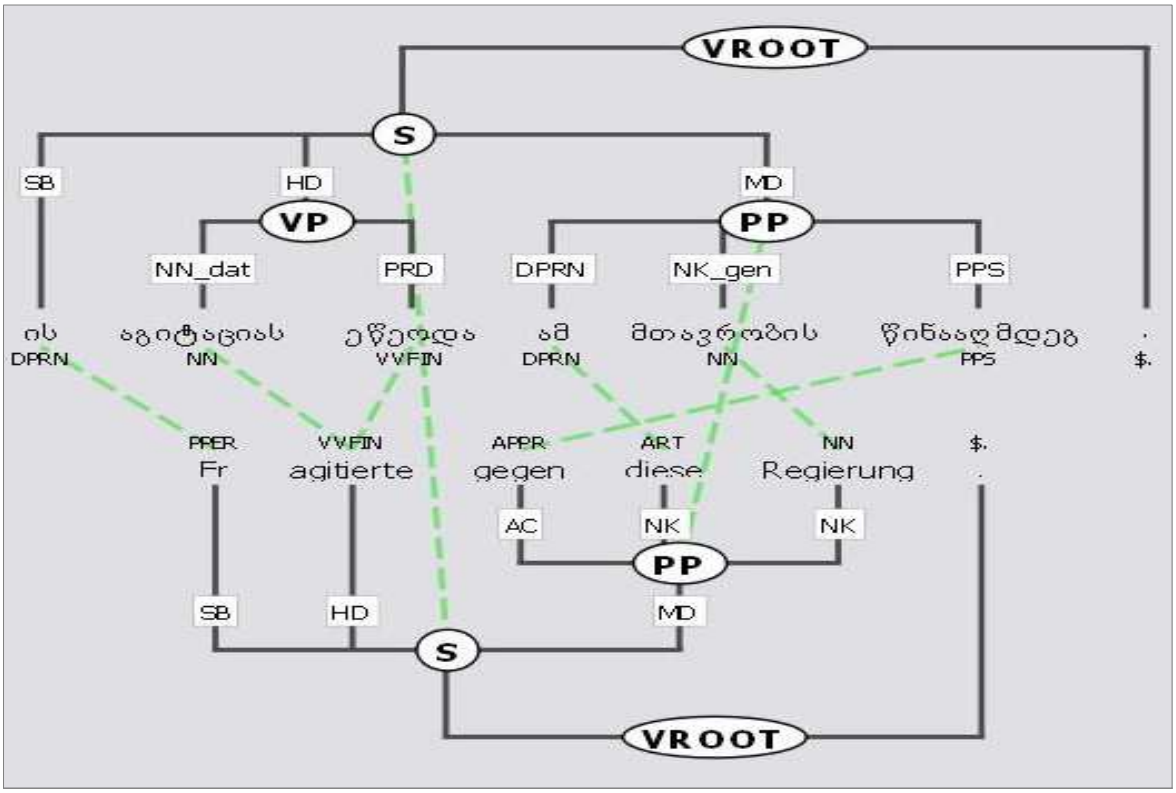

Figure 6: A screenshot with aligned trees from Figure: 1 and Figure: 2.

The Stockholm TreeAligner differentiates between two types of alignment, displayed by different colours. Nodes and words representing exactly the same meaning are aligned as exact translation correspondences using the green colour for lines as it is shown on the Figure 6. In this regard a German word (“agitierte”) alignment to the Georgian Verb Phrase **“აგიტაციის გაწევა”** as an exact one, might be considered problematic.

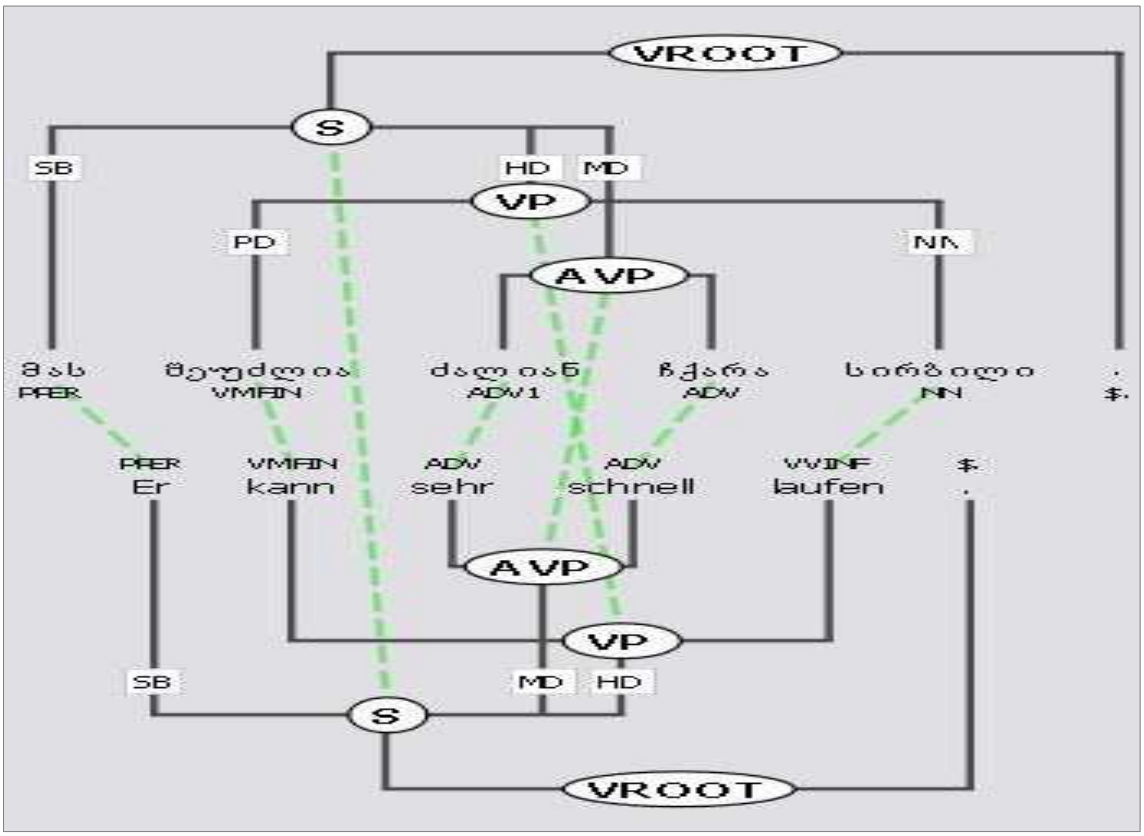

Figure 7: A screenshot of the TreeAligner of the Georgian and German sentences with 1:1 aligned words and phrases.

Nevertheless, in such a case a prerequisite for this solution is that they could serve as translation units outside the current sentence context. If nodes and words represent just approximately the same meaning, they are aligned as fuzzy translation correspondences by means of lines in the red colour as it is shown in the Figure 7 above.

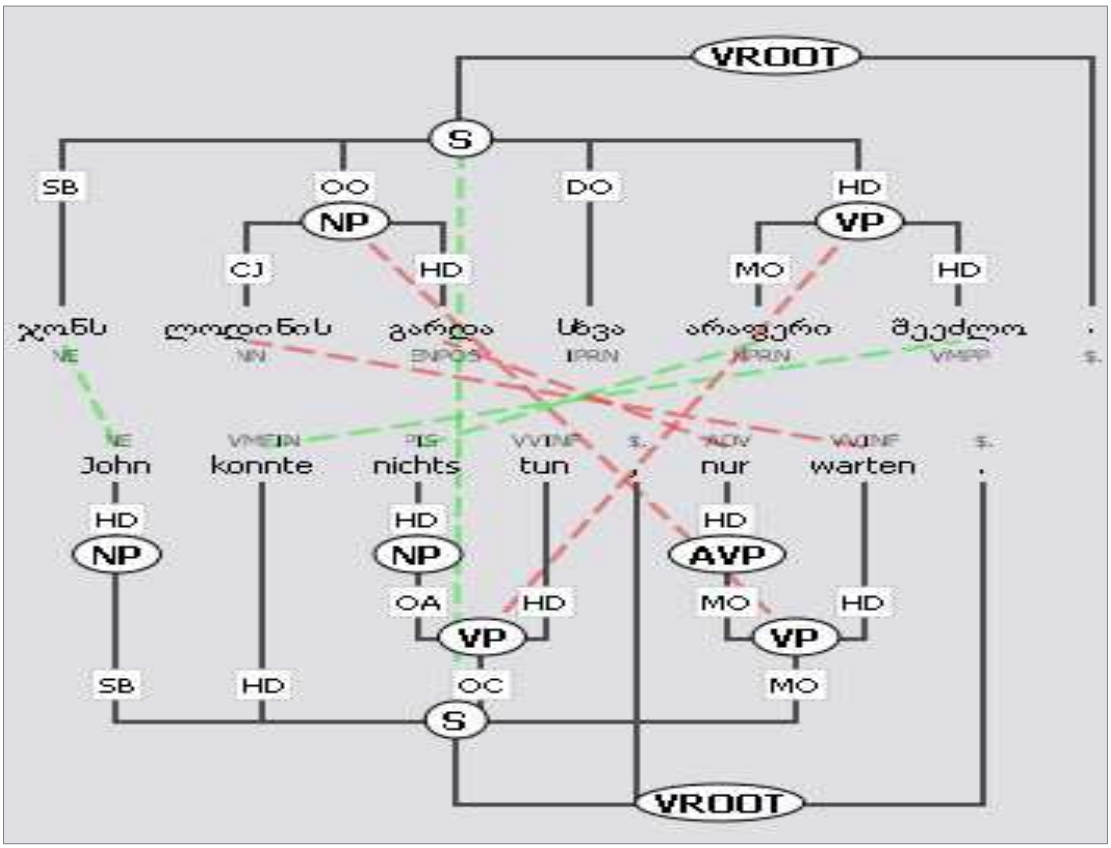

Figure 8: A screenshot of the TreeAligner of the Georgian and German sentences with exact and fuzzy aligned words and phrases.

In an appendix an example of an annotated compound Georgian and German sentences with exact and fuzzy alignment on simple clause and phrase level could be viewed.

5. Conclusions

At the initial phase of presented experiment we made an overview of experience in building parallel threebanks for languages with different structures (Megyesi and Dahlqvist, 2007), (Megyesi et al., 2006), (Grimes et al., 2011), (Rios et al., 2009).

As it is reported in a Quechua-Spanish parallel treebank project, due to strong agglutinative structure of the Quechua language, it was decided to annotate the Quechua treebank on morphemes rather than words. This allowed the authors to link morpho-syntactic information precisely to its source. Besides, building phrase structure trees over Quechua sentences does not capture the characteristics of the language. Therefore, they have chosen Role and Reference Grammar. By using nodes, edges and secondary edges in the Stockholm annotation tool they were able to represent the most important aspects of Role and Reference syntax for Quechua sentences (Rios et al. 2009).

Although the Georgian language is also an agglutinative language with suffixing and prefixing, there is no need to annotate the Georgian Treebank on morphemes. However, for syntactic annotation in the Georgian language a precise description of a specific structure/mechanism of its clause is necessary. “The Georgian clause is a word collocation which draws on coordination and government of the linked verb and noun sequence” [Chikobava, 1928]. The types of

syntactic relations in the Georgian clause differ significantly from those observed in the Indo-European or in other languages. In the English Language there are just a small number of verbs that govern the nouns linked to them as indirect actants and demand those nouns to stand in an indirect case form (e.g. John believes *him* to be innocent). Besides, the actants involved do not induce changes in the verb form. In contrary, in the polyvalent Georgian verb the actants are marked with specific affixes in a verb. The most significant difference from the structure of the Indo-European syntactic relations model is that in the Georgian clause there is a mutual government and agreement relations or a bilateral coordination phenomenon between verb-predicate and noun-actants which number may reach up to three in a single clause. It anticipates control of the noun case forms by verbs, whereas the verbs, in their turn, are governed by nouns with respect to a grammatical person. Therefore, according to [Chikobava, 1928] in a syntactic description of Georgian the concepts of a Major and a Minor Coordinate, instead of Subject and Object, are preferable. Moreover, in the verb forms of a certain semantic type an indirect object has preference as a Major Coordinate over a Subject (a Minor coordinate) in the respect of its marking in a verb form. Nevertheless, unlike the Quechua language, Georgian syntax can be sufficiently well represent by means of dependency relations and there is no need to utilize a different approach to capture the Georgian language structural peculiarities.

The Russian and Ukrainian languages typologically are more closely related languages to German than Georgian is. Consequently, tag-sets for these two languages underwent minor changes and some additional POS and CAT features has been introduced. The changes for the Georgian language tag-sets and CAT values are more significant, but in general they conform to the TIGER guidelines which served as a background in compiling the features and their values for all three new languages involved in the project.

Besides, the Georgian, Russian and German languages also fairly good conform to the TIGER-xml format and syntactic trees perfectly reflect skeletal parses for each those languages.

The TIGER-XML format (.tig extension) is the treebank exchange format allowing free data exchange and the use of tools developed by the international TIGER project community. In the TIGER format, edge labels contain the original syntactic function tags, and the (non-terminal) cat category contains phrase and clause forms. A TIGER-XML file consists of a header and a body. The corpus header can contain meta-information about the corpus (such as corpus name, date, author, etc) and a declaration of the tags that are used in the morphology Part-of-Speech, non-terminal nodes and edges. In the second part of a TIGER-XML file the corpus body contains words, Part-of-Speech tags, morphological tags and lemmata which are listed as attributes of the element "terminal". Non-terminals are represented in an additional element called "nonterminal" referring to the corresponding terminal ID. This part of the XML file contains the encoding for secondary edges as well.

A big advantage of using the xml format is exchangability and usability with a large range of other applications. For example the TIGERsearch corpus query tool, and in the multimedia annotation tools as ELAN and ANNEX.

Monolingual resources with .tig extension for browsing by the Synpathy tool are in the appended folders as followes:

GER – for sentences in the German language;
GEO – for sentences in the Georgian language;
RUS – for the sentences the Russian language;
UKR– for the sentences the Ukrainian language.

Bilingual aligned sentences in .xml format for browsing by means of the Stockholm TreeAligner are in the following folders:

AGEGO – German-Georgian;
AGERU – German- Russian;
AGEUK – German- Ukrainian.

ADDENDUM:

1. Examples of aligned German-Georgia, German-Russian and German-Ukrainian.

2. A presentation for a colloquium in an appended ppt file: ger-geo Treebank.

References

Beesley K. R. and L. Karttunen. (2003). Finite State Morphology. CSLI Publications.

ჩიქობავა, ა. (1928). მარტივი წინადადების პრობლემა ქართულში, თბილისი. [Chikobava A. (1928). The Problem of the Simple Sentence in Georgian. Tbilisi].

Grimes S., Li, X., Bies A., Kulick S. Ma, X. and S. Strassel. (2011). Creating Arabic-English Parallel Word-Aligned Treebank Corpora at LDC. Proceedings of the Second Workshop on Annotation and Exploitation of Parallel Corpora. The 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011). Hissar, Bulgaria. 2011.

Kapanadze O., Kapanadze N., Wanner L., and St. Klatt. (2002). Towards A Semantically Motivated Organization of A Valency Lexicon for Natural Language Processing: A GREG Proposal. Proceedings of the EURALEX conference, Copenhagen.

Kapanadze O. (2010a). Verbal Valency in Multilingual Lexica. In: Workshop Abstracts of the 7th Language Resources and Evaluation Conference-LREC2010. Valletta, Malta.

Kapanadze O. (2010b). Describing Georgian Morphology with a Finite-State System. In A. Yli-Jura et al. (Eds.): Finite-State Methods and Natural Language Processing 2009, Lecture Notes in Artificial Intelligence, Volume 6062, pp.114-122, Springer-Verlag, Berlin Heidelberg .

Kapanadze O. (2009). Finite State Morphology for the Low-Density Georgian Language. In: FSMNLP 2009 Pre-proceedings of the Eighth International Workshop on Finite-State Methods and Natural Language Processing. Pretoria, South Africa

Killer M., Sennrich R. and M. Volk (2011). From Multilingual Web-Archives to Parallel Treebanks in Five Minutes. In Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2011).

Megyesi B. and B. Dahlqvist. (2007). A Turkish-Swedish Parallel Corpus and Tools for its Creation. In Proceedings of Nordiska Datalingvistdagarna (NoDaL- iDa 2007).

Megyesi B., Hein A.S. and E. C. Johanson. (2006). Building a Swedish-Turkish Parallel Corpus. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006).

Rios A., Göhring A. and M. Volk. (2009). Quechua-Spanish Parallel Treebank. In: 7th Conference on Treebanks and Linguistic Theories, Groningen, 2009.

Samuelsson Y. and M. Volk. (2005). Presentation and Representation of Parallel Treebanks. In Proceedings of the Treebank-Workshop at Nodalida, Joensuu, Finland.

Samuelsson Y. and M. Volk. (2006). Phrase Alignment in Parallel Treebanks. In Proceedings of 5th Workshop on Treebanks and Linguistic Theories, Prague, Czech Republic.

Samuelsson Y. and M. Volk. (2007). Alignment Tools for Parallel Treebanks. In GLDV Frühjahrstagung, Tübingen, Germany, 2007.