Fachrichtung 4.6
Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen
Universität des Saarlandes

A German-Georgian Treebank Project

Oleg KAPANADZE
ok@caucasus.net

. This presentation reports about efforts on building a parallel treebank for a typologically dissimilar languages - German and Georgian

. The objective of the project is development of a bilingual Treebank based on a linguistically annotated and syntactically parsed German-Georgian parallel text corpus.

. *Parallel corpora* are language resources that contain texts and their translations, where the texts, paragraphs, sentences and words are linked to each other.

. 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.*

Data for experiment

. For the low-density languages, including Georgian, parallel texts are very rare.

. The parallel corpus used for this study comprises German sentences and their translations into Georgian language compiled for the Georgian-Russian-English-German (GREG) project (Kapanadze et al., 2002, Kapanadze, 2010).

. The GREG lexicon itself contains valency data with the manually aligned Georgian, Russian, English and German verbs (ca. 1250) augmented with examples of sentences considered as translation equivalents.

. A subcorpus used for the experiment has a size more than 2620 sentence pairs that correspond to different syntactic subcategorization frames considered as German-Georgian translation equivalents.

A sketch of the Georgian Grammar

1. Morphology

. Georgian is an agglutinative language using both suffixing and prefixing. Its structure differs from languages of other families such as Indo-European, Semitic or Turkic.

.The noun wordform’s structure in Georgian is as follows:

NOUN_Stem	+	PLURAL_MRK	+	CASE_MRK	+	Emp_V	+	POSTFIX	+	Emp_V
R	+	eb ~ n/t	+	7 options	+	a	+	9 options	+	a

The structural units introduced in red colour are manatory.

There are 7 cases in Georgian with different affixes as the CASE_MRKER and 9 POSTFIXes. Emp_V stands for “an Emphatic Vocal -a”.

. The Georgian verbal patterns are considerably more complex than those of nouns, especially compared to most of the Indo-European languages.

Five classes of verbs are distinguished:

- Transitive verbs
- Intransitive verbs
- Medial verbs
- Inversion verbs
- Stative verbs.

. Rather than using the terms “tense”, “aspect”, “mood”, etc. separately, the Georgian verb grammar is built according to a “*morpho-syntactic*” principle around a construct called *Serie* that is described using the concept of *Screeve*.

. There are 3 Series established according to the syntactic features of Subject or Subject/ Object relations reflected in a verb form. Each screeve is marked for a constant set of tense, aspect and mode.

. A *screeve* is a set of cells in the verbal paradigm, one cell for each *Subjec/Object person/number* combination. The number of sells in a screeve depends upon a valency of the specific verb form.

. If a verb is (intransitive) *monovalent*, then a screeve is a set of six cells in the verbal paradigm, one cell for each subject person/ number combination (subjact1/subject2/ subject3, singular/plural).

. If the verb is *bi- or threevalent* (intransitive or transitive), then the screeve is a set of 22 sells reflecting *Person* and *Subject / Object* relations marked in a verb with the attached syntactic frames (or argument structure such as Subject + Object).

.There are 11 screeves spread across 3 Series.

-An *active* transitive verb root in all 3 Series can produce 242 finite verb forms some of which are ambiguous in the respect of *Person* and *Subject ~ Object* relations.

-The majority of the *active* transitive verb roots can be converted and inflected as intransitive *bivalent* or *monovalent* verbs.

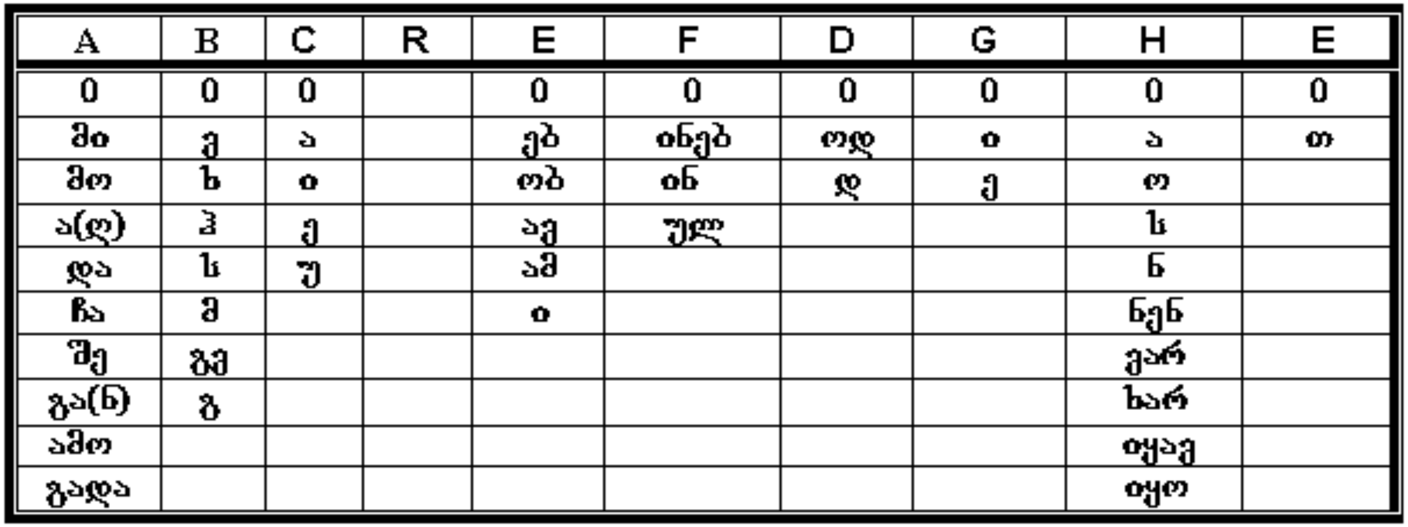
-Theoretically, a single Georgian verb root is capable to produce more than 1000 different finite verb forms.

- Each verb screeve is formed by adding a number of prefixes and suffixes to a verb root.

-Certain affix categories are limited to certain screeves. In a given screeve, not all possible markers are obligatory.

-The overall structure can be visualized as linear sequence of positions, or ‘slots’, before and after the root position, which is referred to as slot *R*:

A+B+C+R+D+E+F+G+H



A	B	C	R	E	F	D	G	H	E
0	0	0		0	0	0	0	0	0
მი	ვ	ა		ებ	ინებ	ოდ	ი	ა	თ
მო	ხ	ი		ობ	ინ	დ	ე	ო	
ა(ღ)	ჰ	ე		ავ	ულ			ს	
და	ს	უ		ამ				ნ	
ჩა	მ			ი				ნენ	
შე	გვ							ვარ	
გა(ნ)	გ							ხარ	
ამო								იყავ	
გადა								იყო	

-In the liner rank order the elements of the *rank* R have the highest priority in the sense of generative constraints. Derived from the verb class they license the most of the elements from the other ranks.

-**A typical result of the Georgian verb analyses contains a morphological structure of a finite verb and its syntactic valency. *e.g.***

vyidiT (*"we sell it/them"*)

Subj1/v + yid + theme/i + T=atsmko/Subj1P1 + Obj3Sg

Subj1/v + yid + theme/i + T=atsmko/Subj1P1 + Obj3Pl

In the example above

-“*v*” reffers the 1st person Subject (*Subj1)*

-“*yid*” is a verbal root

-“*i*” is a thematic marker, an element that forms a steme for the first Serie verb paradigme.

-“*T*” is a plural marker for the 1st person Subject .

2. Syntax

- The types of syntactic relations in the Georgian clause differ significantly from that observed in the Indo-European or in other languages.
- “The Georgian clause is a word collocation which draws on coordination and government of the linked verb and noun sequence” [Chikobava, 1928].

.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 in a clause 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 Indo-European syntactic relations model is that in the Georgian clause we have a mutual government and agreement relations or a bilateral coordination 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.

. 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.

.

Building Monolingual Treebanks

We tried to experiment with several software packages:

- Geometric Mapping and Alignment -GMA
  (I. Dan Melamed. Empirical Methods for Exploiting Parallel Texts. 2001. MIT Press.)
- Champollion Tool Kit V1.1
  (Xiaoyi Ma, Linguistic Data Consortium)
- Annotate (Oliver Plaehn, CL, Univ. of Saarland)
- TreeEditor
- Synpathy (Max-Plannk Institute for Psycholinguistics, Nijmagen

Opting for Synpathy as syntax editor we made an overview of experience in building parallel threebanks for languages with different structures:

- **Turkish-Swedish (Megyesi and Dahlqvist, 2007), (Megyesi et al., 2006),**

- **Arabic-English (Grimes et al., 2011),**

- **Quechua-Spanish (Rios et al., 2009).**

. 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.

. Reportedly, building phrase structure trees over Quechua sentences does not capture the characteristics of the language. Therefore, for Quechua annotation was chosen Role and Reference Grammar.

. Although the Georgian language is also an agglutinative language with suffixing and prefixing, there is no need to annotate a monolingual Georgian Treebank on morphemes. Besides, unlike the Quechua language, Georgian syntax can be sufficiently well represent by means of dependency relations.

. For morphological analyses of Georgian sentences we applied a finite-state 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.

. 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.

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.

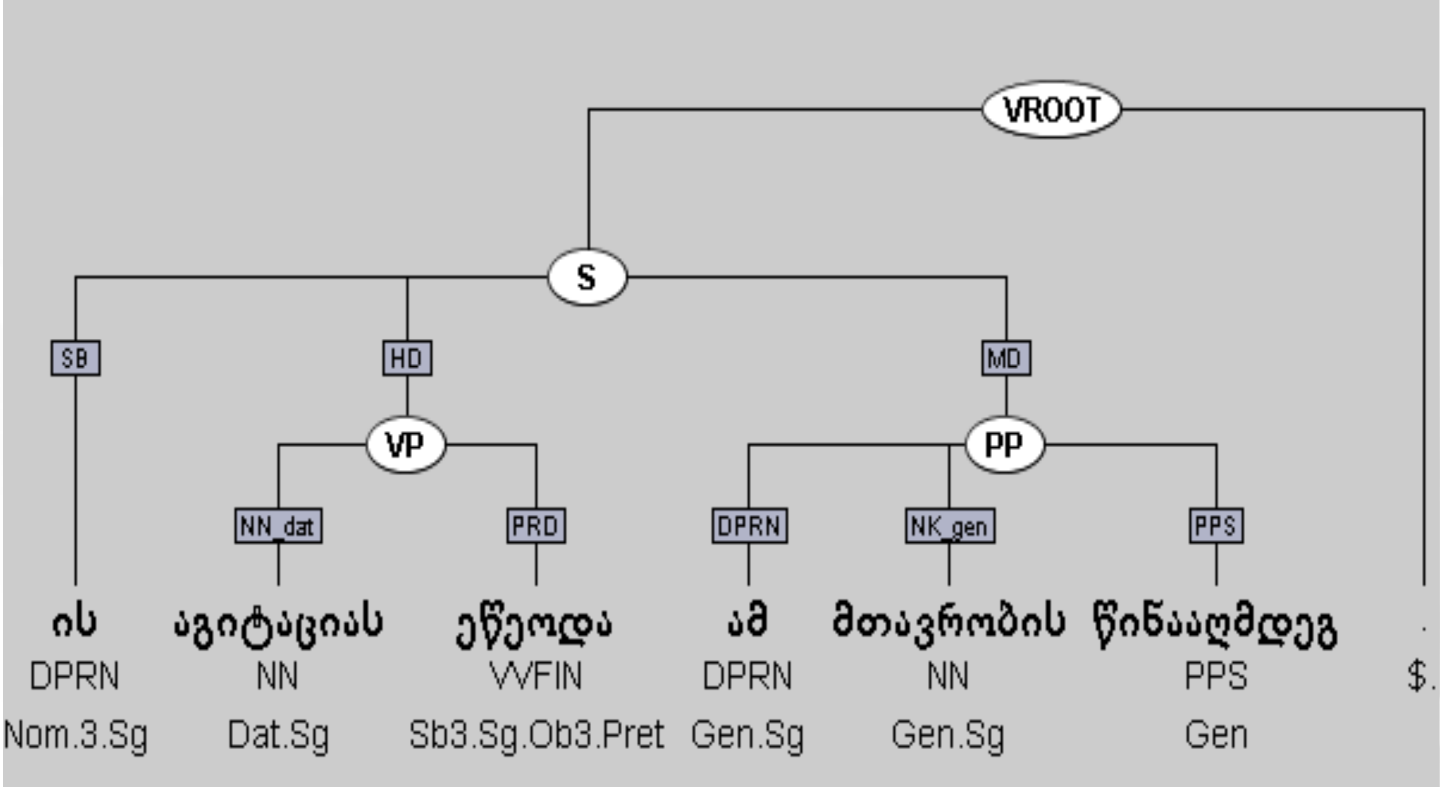
. The morphologically annotated bilingual German and Georgian corpus has been parsed using *the Synpathy,* a tool for syntactic annotation.

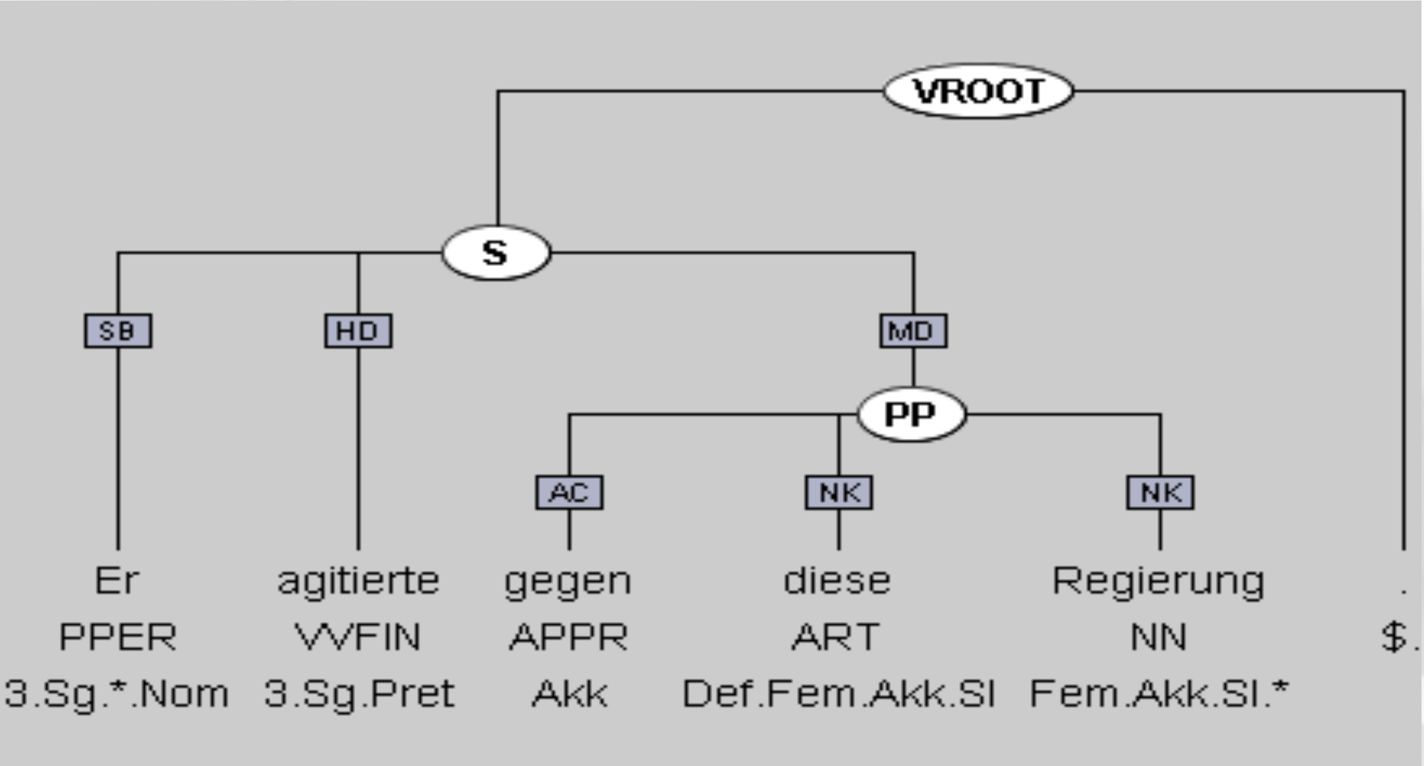
. The German treebank annotation follows the TIGER annotation scheme (Skut et al., 1997, Brants et al., 2002).

. The Georgian treebank was annotated according an adapted version of the TIGER guidelines with the necessary changes relevant to the Georgian grammar.

.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 German and in Georgian.

.The monolingual treebanks converted into TIGER-XML, are a powerful database-oriented representation for graph structures.


VROOT
S
SB
HD
MD
VP
PP
NN_dat
PRD
DPRN
NK_gen
PPS
ის
DPRN
Nom.3.Sg
აგიტაციას
NN
Dat.Sg
ეწეოდა
VVFIN
Sb3.Sg.Ob3.Pret
ამ
DPRN
Gen.Sg
მთავრობის
NN
Gen.Sg
წინააღმდეგ
PPS
Gen
.
$.


VROOT
S
SB
HD
MD
PP
AC
NK
NK
Er
PPER
3.Sg.*.Nom
agitierte
VVFIN
3.Sg.Pret
gegen
APPR
Akk
diese
ART
Def.Fem.Akk.Sl
Regierung
NN
Fem.Akk.Sl.*
.
$.

. In a TIGER-XML graph each leaf (= token) and each node (= linguistic constituent) has a unique identifier. 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.

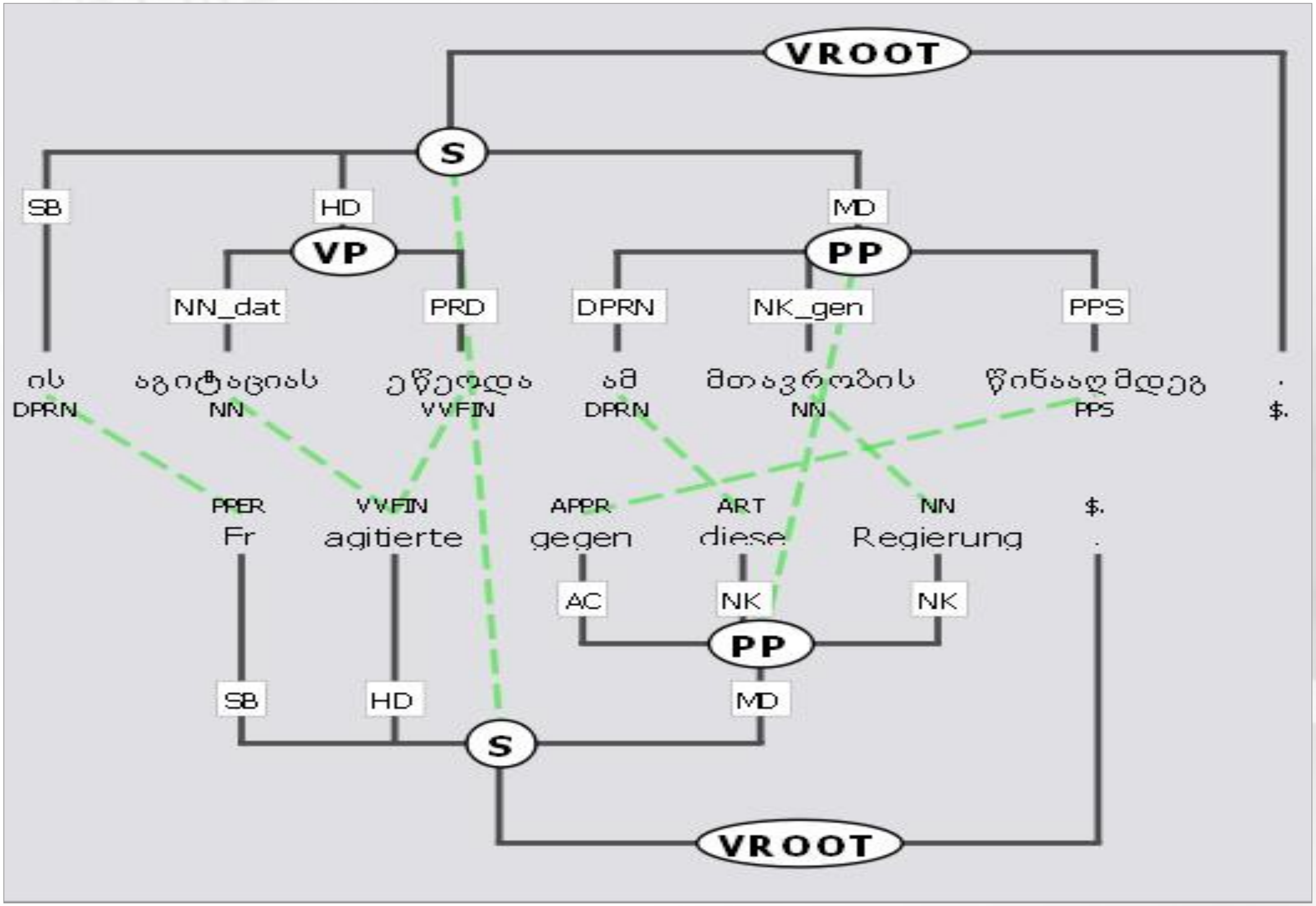
Building a Parallel Treebank from a Monolingual German and a Monolingual Georgian Treebanks

. The alignment procedure is done by means of *the Stockholm TreeAligner,* a tool for work with parallel treebanks which inserts alignments between pairs of syntax trees.

. 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 Georgian language. This is done with the help of a graphical user interface of the Stockholm TreeAligner.

.The alignment lines are drawn manually between pairs of sentences, phrases and words over parallel syntax trees.


VROOT
S
SB
HD
MD
VP
PP
NN_dat
PRD
DPRN
NK_gen
PPS
ის
DPRN
აგიტაციას
NN
ეწეოდა
VVFIN
ამ
DPRN
მთავრობის
NN
წინააღმდეგ
PPS
.
$.
PPER
Er
VVFIN
agitierte
APPR
gegen
ART
diese
NN
Regierung
$.
.
AC
NK
NK
PP
SB
HD
MD
S
VROOT

. In a TIGER-XML graph each leaf (= token) and each node (= linguistic constituent) has a unique identifier. 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.

We intended to align as many phrases as possible. The goal is to show translation equivalents. 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.

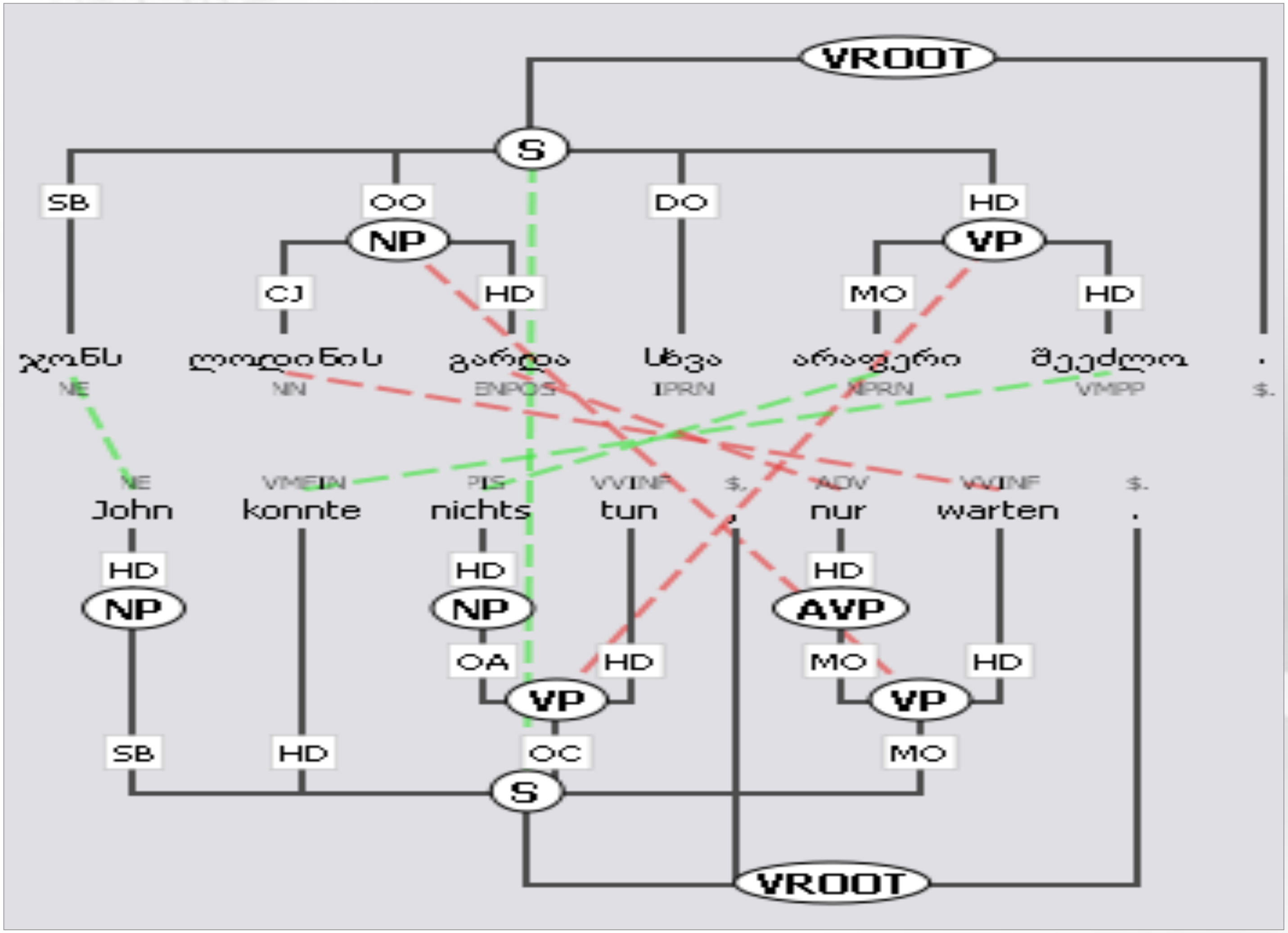
-One example of 1: n alignment on the word level is the Georgian multi-word expression for “აგიტაციის გაწევა” represented under a VP node, which is one word (“agitierte”) in the corresponding German sentence.

-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.

- 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.

-In this regard a German word (“agitierte”) alignment to the Georgian Verb Phrase “აგიტაციის გაწევა” as an exact one, might be considered problematic. 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.


VROOT
S
SB
OO
DO
HD
NP
VP
CJ
HD
MO
HD
ჯონს
NE
ლოდინის
NN
გარდა
BNPOS
სხვა
IPRN
არაფერი
XPRN
შეეძლო
VMPP
.
$.
NE
John
VMFIN
konnte
PIS
nichts
VVINF
tun
$,
,
ADV
nur
VVINF
warten
$.
.
HD
NP
HD
NP
HD
AVP
OA
HD
MO
HD
VP
VP
SB
HD
OC
MO
S
VROOT

Future Research

Parallel texts for the Georgian language are very rare. Nevertheless, recently, we have discovered a bilingual repository comprising the German-Georgian texts in jurisprudence.

This bilingual text corpus is a collection of the Georgian laws translation into the German language created by human translators thanks to the GTZ (*Gesellschaft für Technische Zusammenarbeit*) and German jurists.

- In the presented project, except morphological analysis for the Georgian text, syntax structures for both languages had been created completely manually.

- For further expending the parallel German-Georgian Treebank for legislative texts we intend to experiment with the Tree-to-Tree (t2t) Alignment Pipe (Killer, Sennrich and Volk, 2011a, b).

t2t is a collection of python scripts, generating automatically aligned parallel treebanks from multilingual web resources or existing parallel corpora. The pipe contains wrappers for a number of freely available NLP software programs used for

- tokenization (NLTK Treebank),
- sentence alignment (Hunalign, Vanila Aligner, Microsoft BSA),
- word alignment (Giza++),
- syntactic parsing (Stanford Parser, Berkley Parser)
- Tree to Tree Aligner (Zhechev 2009).

On the final stage, the Tree-to-Tree alignment pipe prepares an input for its further processing in the Graphical User Interface of the Stockholm TreeAlligner which will be used for the aligned German-Georgian Treebank visualization and correction procedures.

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. (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.

Zhechev V. (2009). Automatic Generation of Parallel Treebanks. An Efficient Unsupervised System. Dissertation.
http://www.ventsislavzhechev.eu/Home/
Publications- files/PhD%20Thesis
%20Zhechev20%E2%80%94%C2%A0Final.pdf.

THANK YOU

For Your Patient.