The Multilingual GRUG Parallel Treebank – Syntactic Annotation for Under-Resourced Languages

Oleg Kapanadze

Ilia State University
Georgia

ok@caucasus.net

TREEBANKS

- This presentation reports about efforts on building a multilingual repository for Under-Resourced languages

- The objective of the project is development of bilingual TreeBanks based on the morphologically annotated and syntactically parsed parallel text corpora

- ***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**

- A variety of languages falls into the Under-Resourced category, from languages, for which linguistic descriptions abound, to those, which do not yet even have a standard lexical stock

- Building annotated corpora and constructing treebanks lead to improvement of grammars and lexicons

- This is especially relevant for the under-resourced languages

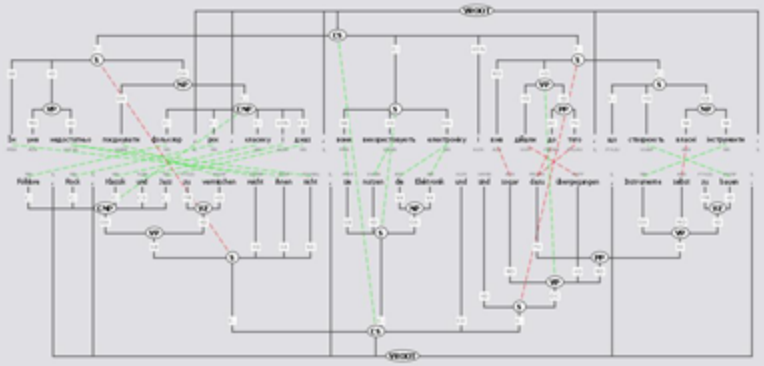
CLARIN-D

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

GRUG parallel treebank

Georgian
Russian
Ukrainian
German



UNIVERSITÄT DES SAARLANDES

Welcome to the website of the GRUG Parallel Treebank

24th October 2012
Georgian-Ukrainian added.

1st October 2012
Website published!

20th August 2012
Draft of the website

27th July 2012
PID / handle issued

27th July 2012
CMDI and DC metac

HOME

Documentation
Downloads
Contact
Metadata
License

Description

This dataset is made of two types of resources: four monolingual Treebanks (German, Georgian, Russian and Ukrainian), and four parallel Treebanks (German-Georgian, German-Russian, German-Ukrainian, Georgian-Ukrainian). 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. 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.

Morphological analysis

GRUG PROJECT

GRUG: Georgian-Russian-Ukrainian-German Treebank

http://fedora.clarin-d.uni-saarland.de/grug/

Monolingual TreeBanks : GE, RU, UK, GO

Under-Resourced languages: GE,UA

Bilingual TreeBanks:

GE-RU

GE-GO

GE-UK

GO-UK

A Georgian, Russian, English and German Valency Lexicon for Natural Language Processing

(INTAS Georgia Project, INTAS 1921)

GREG was an INTAS Georgia project between the Intelligent Systems Group of the Computer Science Department at the University of Stuttgart (coordinator), the Information Technology Research Institute at the University of Brighton, the Computational Lexicography Group, Department of Applied Mathematics and Computer Science at the Tbilisi State University, and the Institute of Linguistics, Georgian Academy of Sciences.

The goal of the GREG project was to develop a multilingual valency lexicon that is suitable for use in various NLP-applications. It contains contain syntactic and semantic valency patterns for about 1000 verbal equivalents from English, Georgian, German and Russian. Syntactic valency patterns are specified in terms of subcategorization frames; semantic valency patterns draw on the thematic (or functional) roles as introduced by Halliday, Chafe, Longacre and others.

Both the micro and the macro structure of the lexicon follow the general principles for an efficient representation of multilingual valency lexica that have also been developed in the project. The representation formalism used to encode the GREG lexicon is DATR.

The project started in March 1999 and terminated in August 2001.

GREG corpus for GRUG

GREG (**German-Russian-English-Georgian**)

A Multilingual Valency Lexicon for NLP extended with examples of sentences in the mentioned four languages

The appended sentences unfold lexical entries' meaning and are considered as mutual translation equivalents of the languages which the GREG lexicon covers

GREG corpus for GRUG

Visualization Material:

English – Tree-adjoined Grammar, ITRI, Brighton

German –Syntactic Subcategorization Trees, Institut für Maschinelle Sprachverarbeitung, U Stuttgart

Russian - Hand crafted subcategorizatrion frames

Georgian - Hand crafted subcategorizatrion frames

Levels of Subcategorization:

Syntactic Valency Frames

Semantic Valency (Deep Cases/Semantic Roles) labels

TASKS in GRUG

- Morphological Annotation and Syntactic Alignment of the monolingual Georgian, Russian, Ukrainian, German text corpora using ***Synpathy software***

- Development of Bilingual Linguistically Annotated TreeBanks Using ***Stockholm TreeAligner (TA)*** graphical interface

- Testing ***Synpathy*** and ***Stockholm TreeAligner*** for non Latin script (Cyrillic and Georgian)corpus.

Steps for GRUG Building

1. Development of Monolingual TreeBanks:

- Compiling Scripts for monolingual Syntactic Parsing of Text in Four Languages (**.tig**)
- Building Monolingual (**GE, RU, UK,GO**) TreeBanks

2. Converting .tig -Format into .xml -Format

3. Alignment of Monolingual TreeBanks into Bilingual Parallel GE-RU, GE-GO, GE-UK, GO-UK TreeBanks

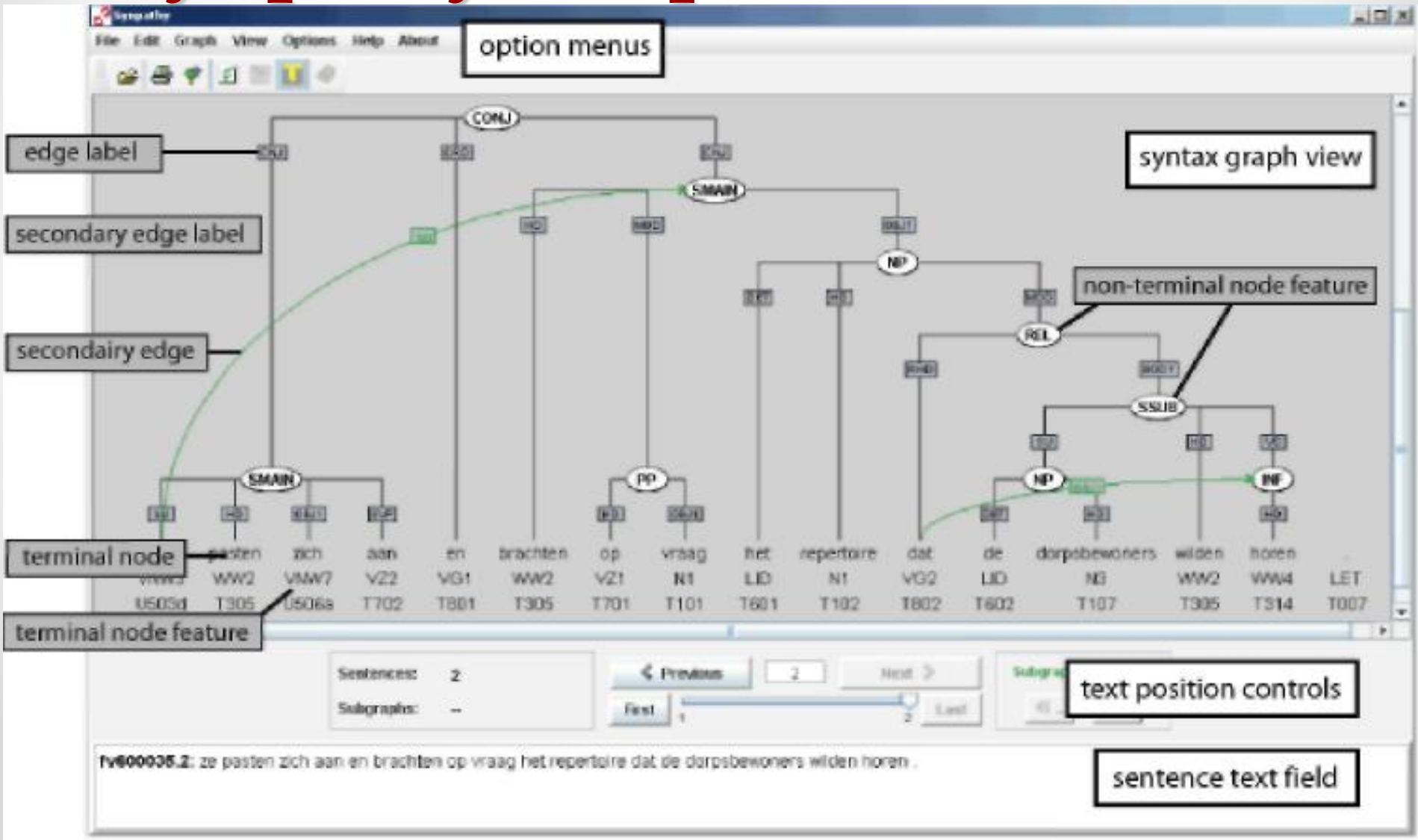
An excerpt of a script for encoding of a sentence

“ჯონს ლოდინის გარდა სხვა არაფერი შეეძლო”

(John Could do nothing, but to wait)

```
<terminals>
  <t id="s692_1" word="ჯონს" pos="NE" morph="Dat.Sg." />
  <t id="s692_2" word="ლოდინის" pos="NN" morph="Gen.Sg." />
  <t id="s692_3" word="გარდა" pos="ENPOS" morph="--" />
  <t id="s692_4" word="სხვა" pos="IPRN" morph="Nom.Sg." />
  <t id="s692_5" word="არაფერი" pos="NPRN" morph="Nom.Sg."/>
  <t id="s692_6" word="შეეძლო" pos="VMFIN" morph="3.Sg.Past.Ind" />
  <t id="s692_7" word="." pos="$." morph="--" />
</terminals>
<nonterminals>
<nt id="s692_502" cat="S">
    <edge label="SB" idref="s692_1"/>
    <edge label="OO" idref="s692_500"/>
    <edge label="DO" idref="s692_4"/>
    <edge label="HD" idref="s692_501"/>
</nt>
<nt id="s692_500" cat="NP">
    <edge label="CJ" idref="s692_2"/>
    <edge label="HD" idref="s692_3"/>
</nt>
<nt id="s692_501" cat="VP">
    <edge label="MO" idref="s692_5"/>
    <edge label="HD" idref="s692_6"/>
</nt>
<nt id="s692_VROOT" cat="VROOT">
  <edge label="--" idref="s692_502" />
  <edge label="--" idref="s692_7" />
```

Synpathy Graphical Interface



Visualization of Monolingual TreeBanks

Synpathy *Grafical Interface,* Windows, MAC OS, Linux:
http://www.mpi.nl/tools/synpathy.html.
(Max Planck Institute for Psycholinguistics, Nijmegen)

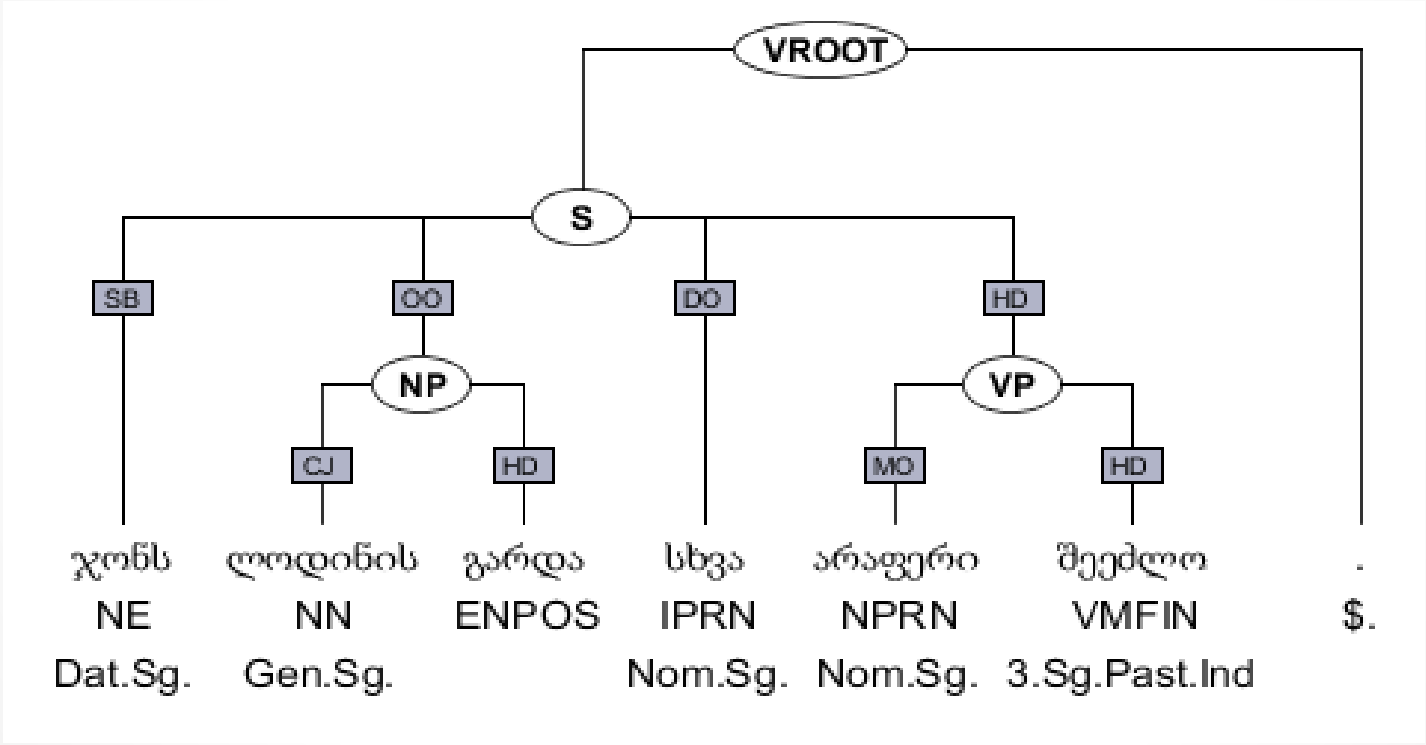
TIGER-Format displays tree-like graph structures:

- the node labels are phrasal categories

- the parental and secondary edge labels contain function labels (subject, object, attribute, etc.)

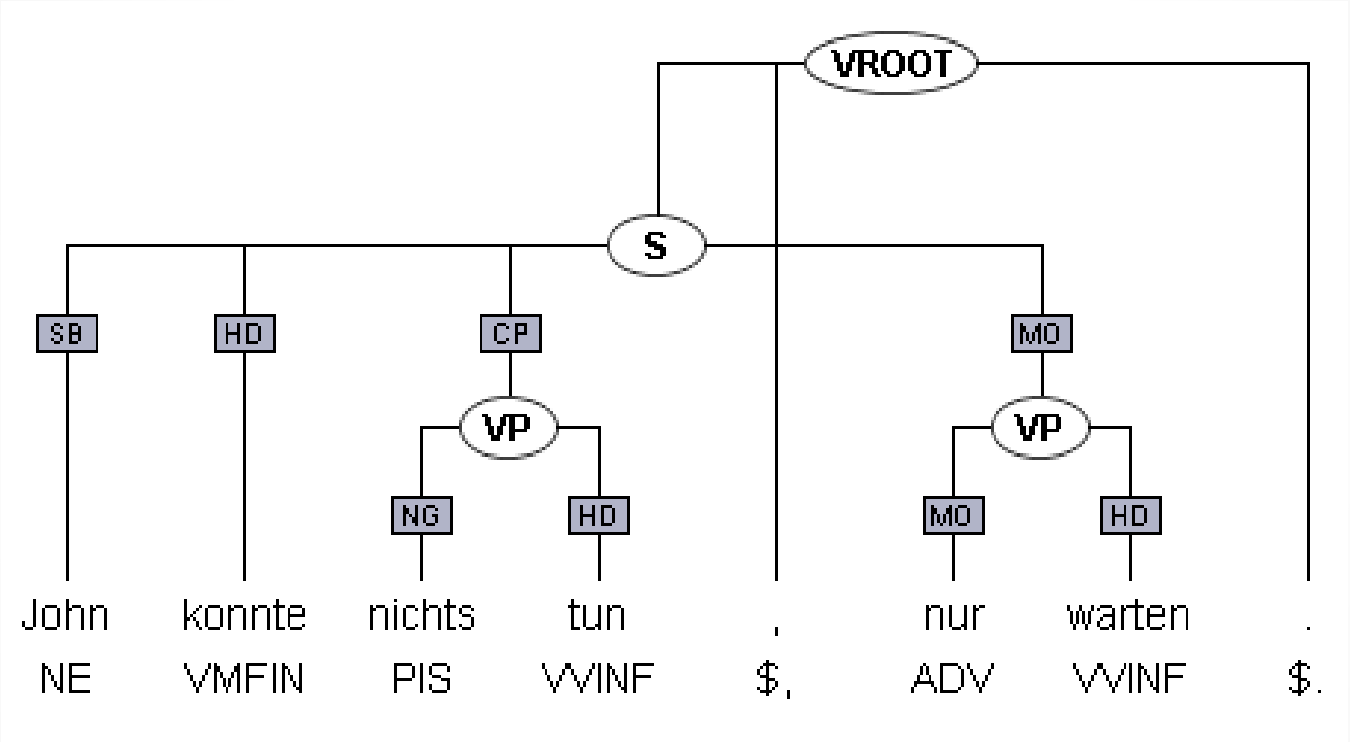
“ჯონს ლოდინის გარდა სხვა არაფერი შეეძლო”.

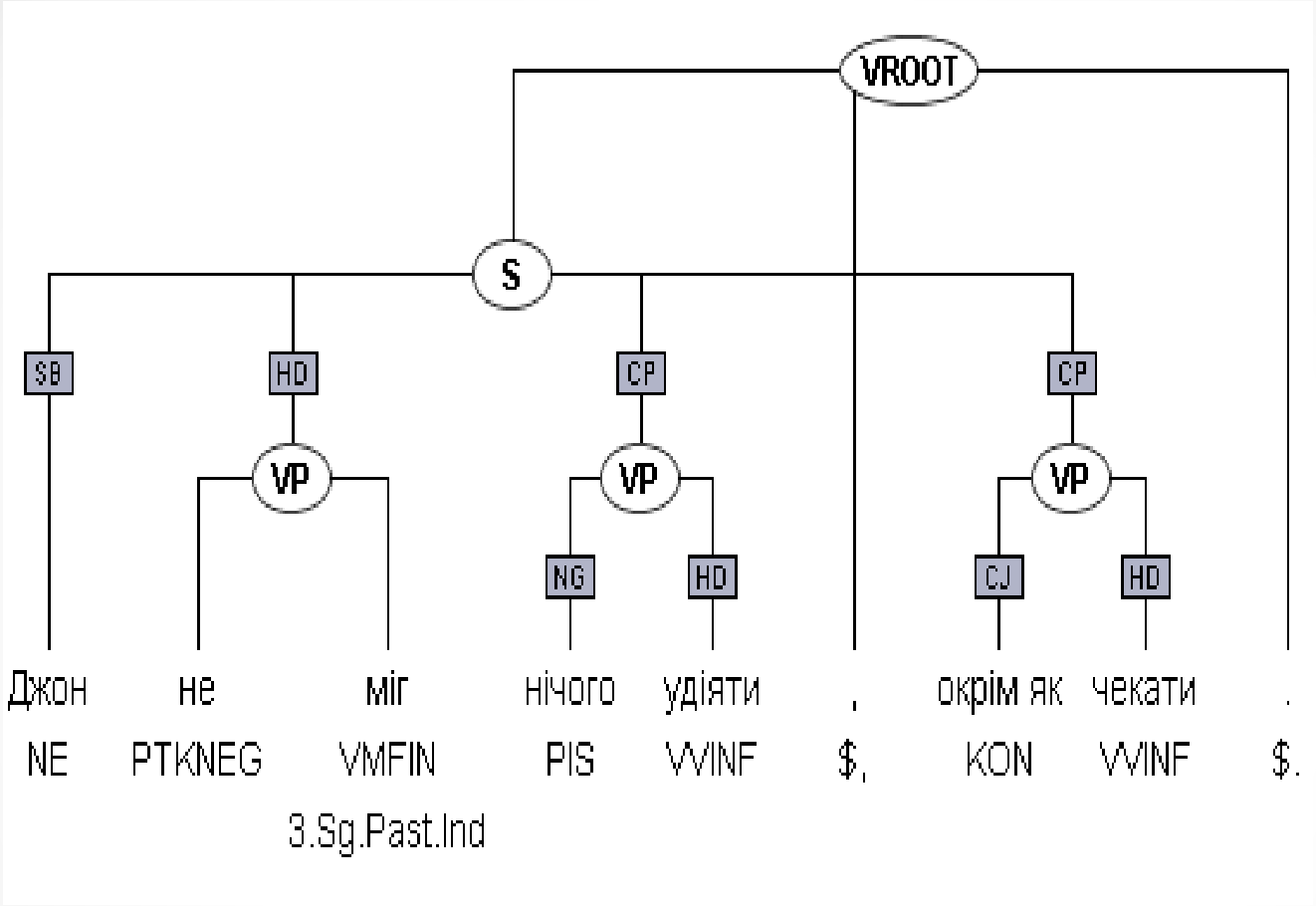
(John could do nothing, but to wait)

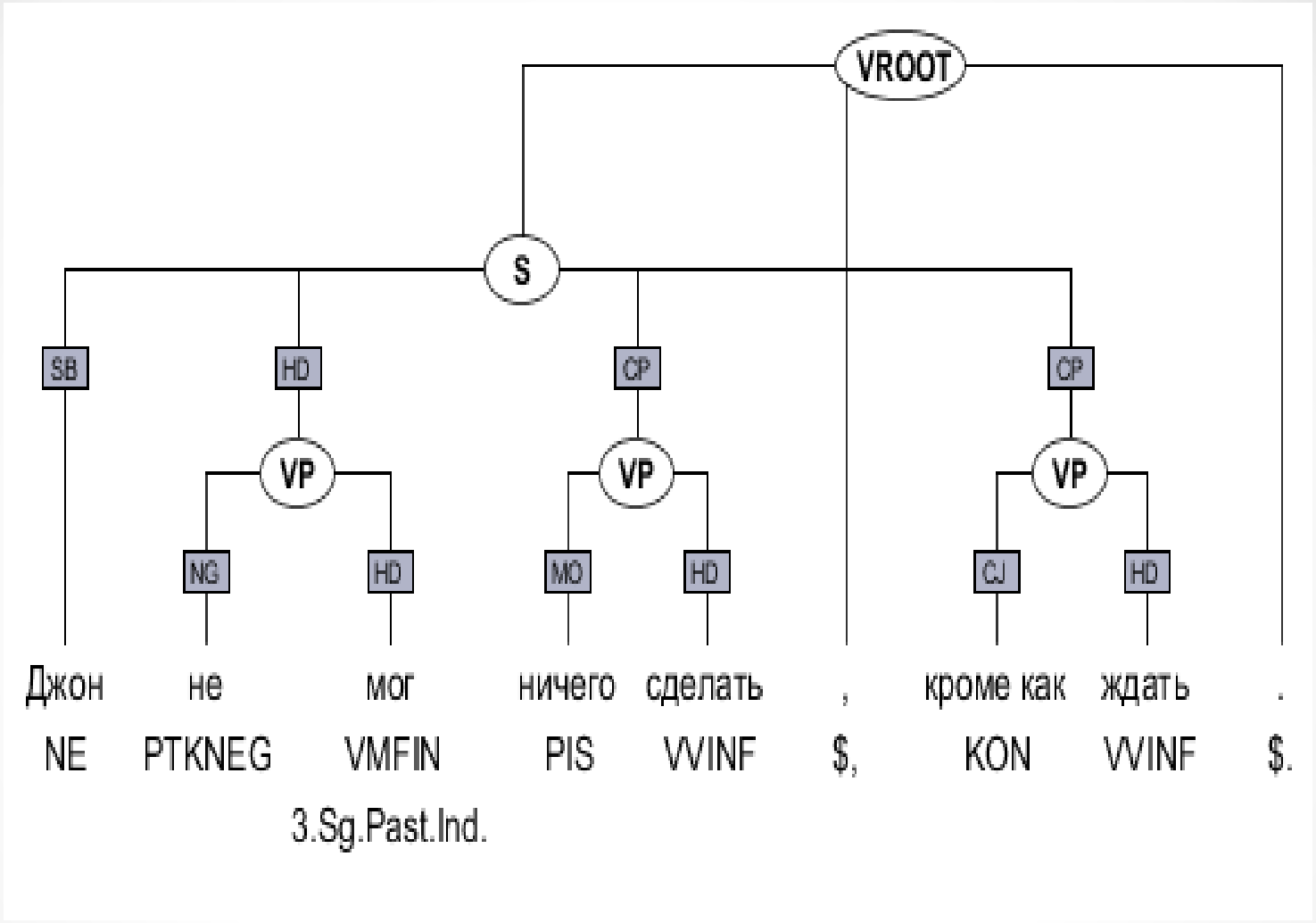


“John konnte nichts tun, nur warten”.

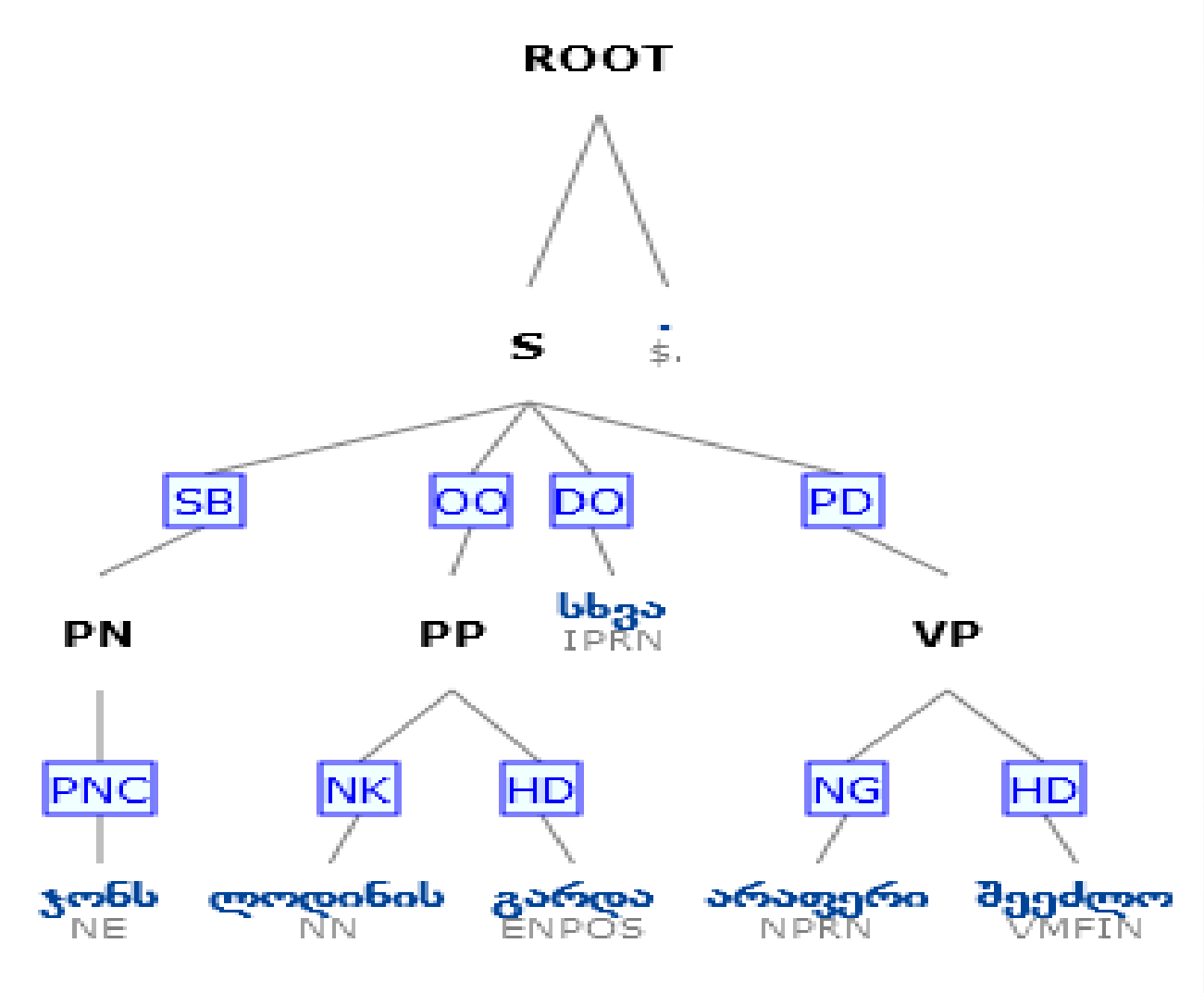
(John could do nothing, but to wait)




VROOT
S
SB
HD
CP
CP
VP
VP
VP
NG
HD
CJ
HD
Джон
не
міг
нічого
удіяти
,
окрім
як
чекати
.
NE
PTKNEG
VMFIN
PIS
VVINF
$,
KON
VVINF
$.
3.Sg.Past.Ind


VROOT
S
SB
HD
CP
CP
VP
VP
VP
NG
HD
MO
HD
CJ
HD
Джон
не
мог
ничего
сделать
,
кроме как
ждать
.
NE
PTKNEG
VMFIN
PIS
VVINF
$,
KON
VVINF
$.
3.Sg.Past.Ind.

INESS Project Option for Visualization



Syntactic Categories

Tag-Sets for **German** follow the NEGRA project STTS (Stuttgart-Tübinger Tagset, G. Smith, 2003)

For **RU, UK, GO** the STTS are expended by the **POS** tags and features necessary to visualize peculiarities of syntactic structures in those languages:

Nouns N
Adverbs ADV
Verbs V
Conjunctions KON
Articles ART
Adpositions AP
Adjectives ADJ
Interjections IT
Pronouns P
Particles PTK
Cardinal Numbers CARD

Features for German Verb

Part of Speech	Typ des Verbes	Finiteness	Beispiele
V	A 'Auxiliar'	FIN 'finit' INF 'infinit' IMP 'imperativ' PP 'Partizip Perfekt'	VVINF 'haben VAIMP 'sein' VAPP 'hatte
	M 'Modal'	FIN 'finit' INF 'infinit' PP 'Partizip Perfekt'	VMFIN 'könnte' VMINF 'können' VMPP 'gekonnt'
	F 'Full'	FIN 'finit' INF 'infinit' IZU 'Infinitiv mit zu' IMP 'imperativ' PP 'Partizip Perfekt'	VFFIN 'gebt … ab' VFINF 'abgeben' VFIZU 'abzugeben' VFIMP 'gib … zu' VFPP 'abzugegeben'

Building the Parallel Treebanks

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

***Synphaty* Output .tig → *STA* Input .xml**

***The Stockholm TreeAligner* handles alignment of tree structures, in addition to word alignment, which – according to its developers - is unique.**

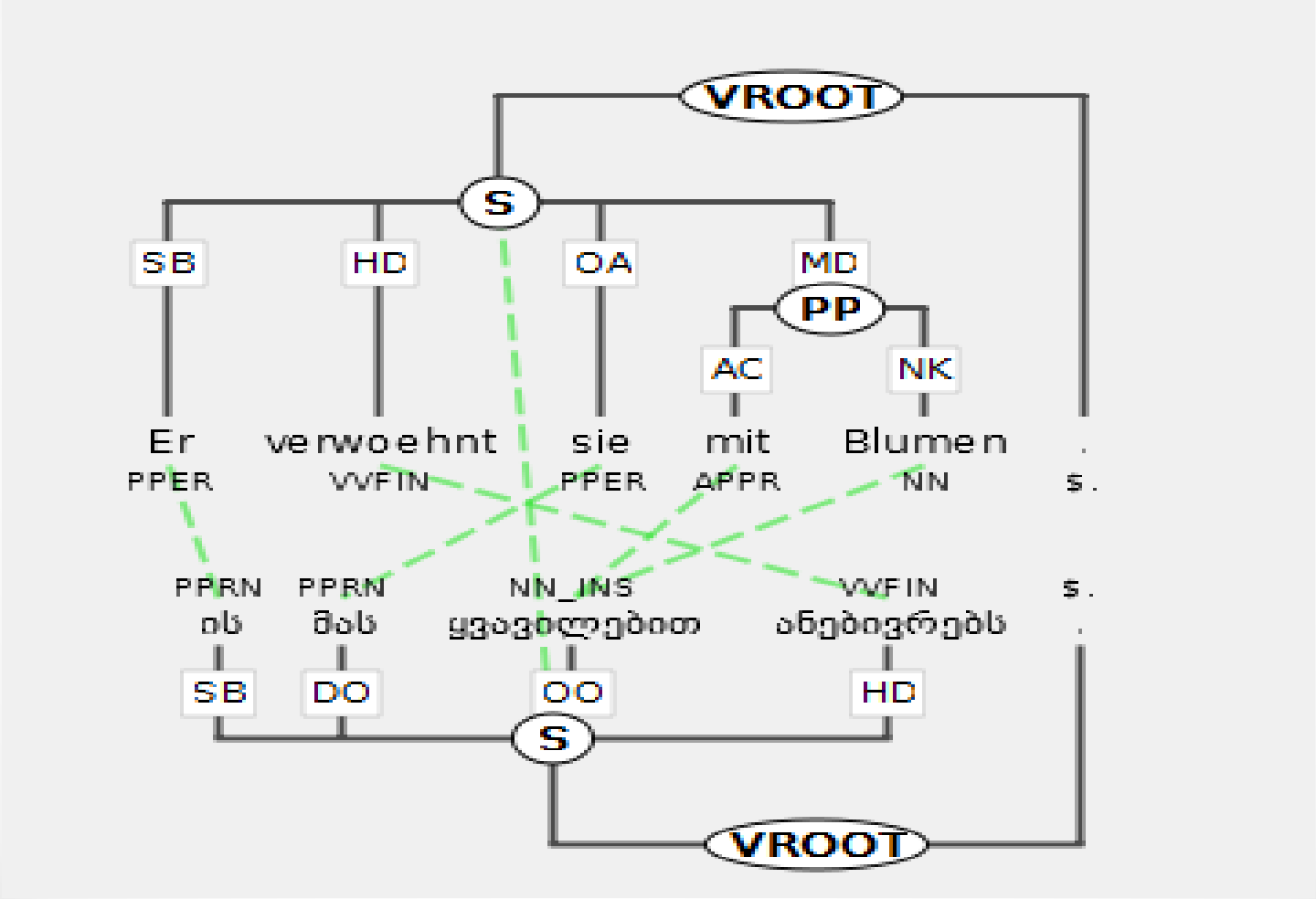
Colours:

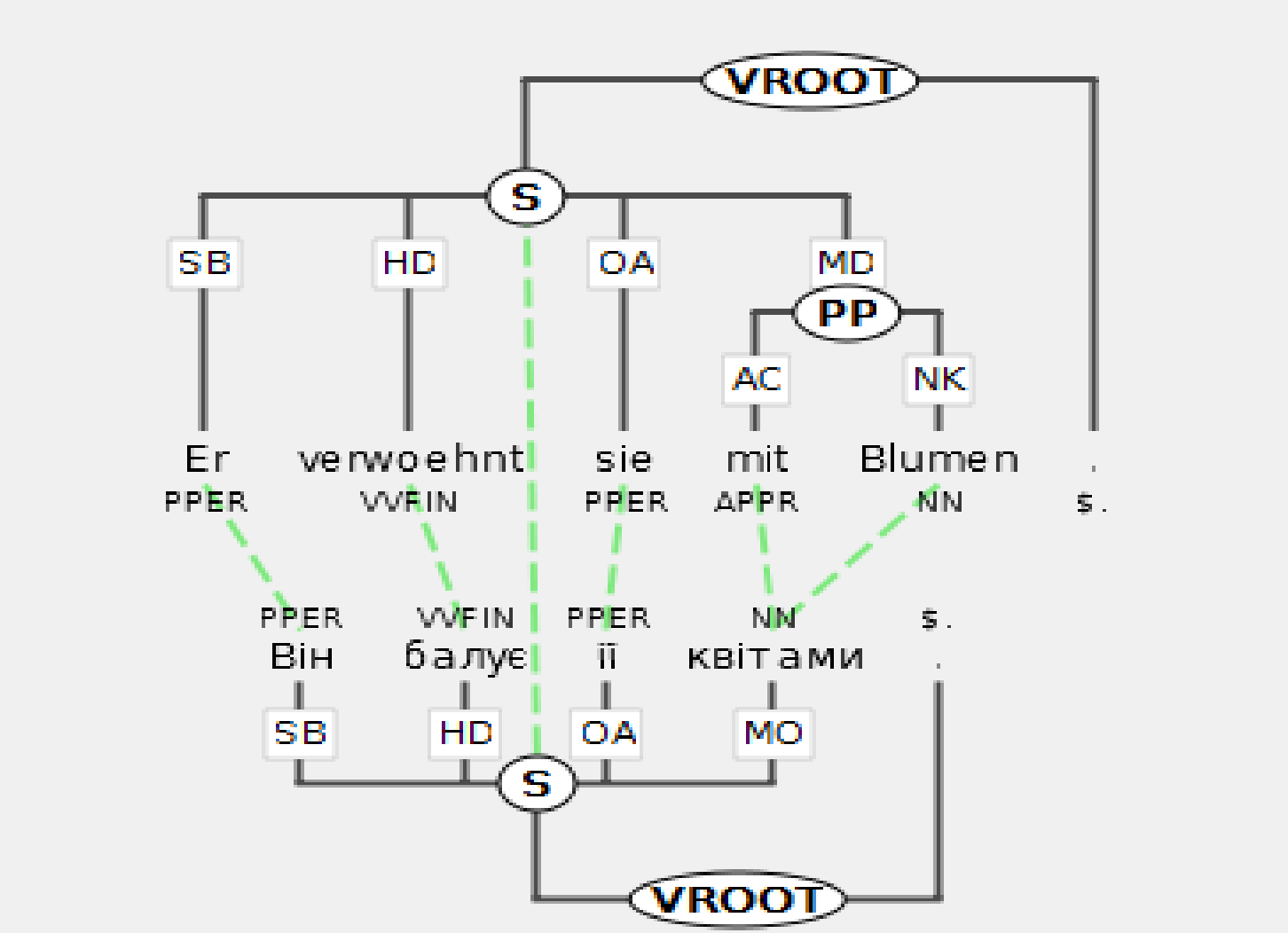
Green – exact matches

Red – fuzzy matches

“Er verwöhnt sie mit Blumen”

(lit. He cossets her with flowers).




VROOT
S
SB
HD
OA
MD
PP
AC
NK
Er
verwoehnt
sie
mit
Blumen
.
PPER
VVFIN
PPER
APPR
NN
$.
PPER
VVFIN
PPER
NN
$.
Він
балує
її
квітами
.
SB
HD
OA
MO
S
VROOT

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

. Phrase alignment is an additional layer of information on top of the syntax structure. It shows which part of a sentence in the L1 language is equivalent to a part of a corresponding sentence in the L2 language

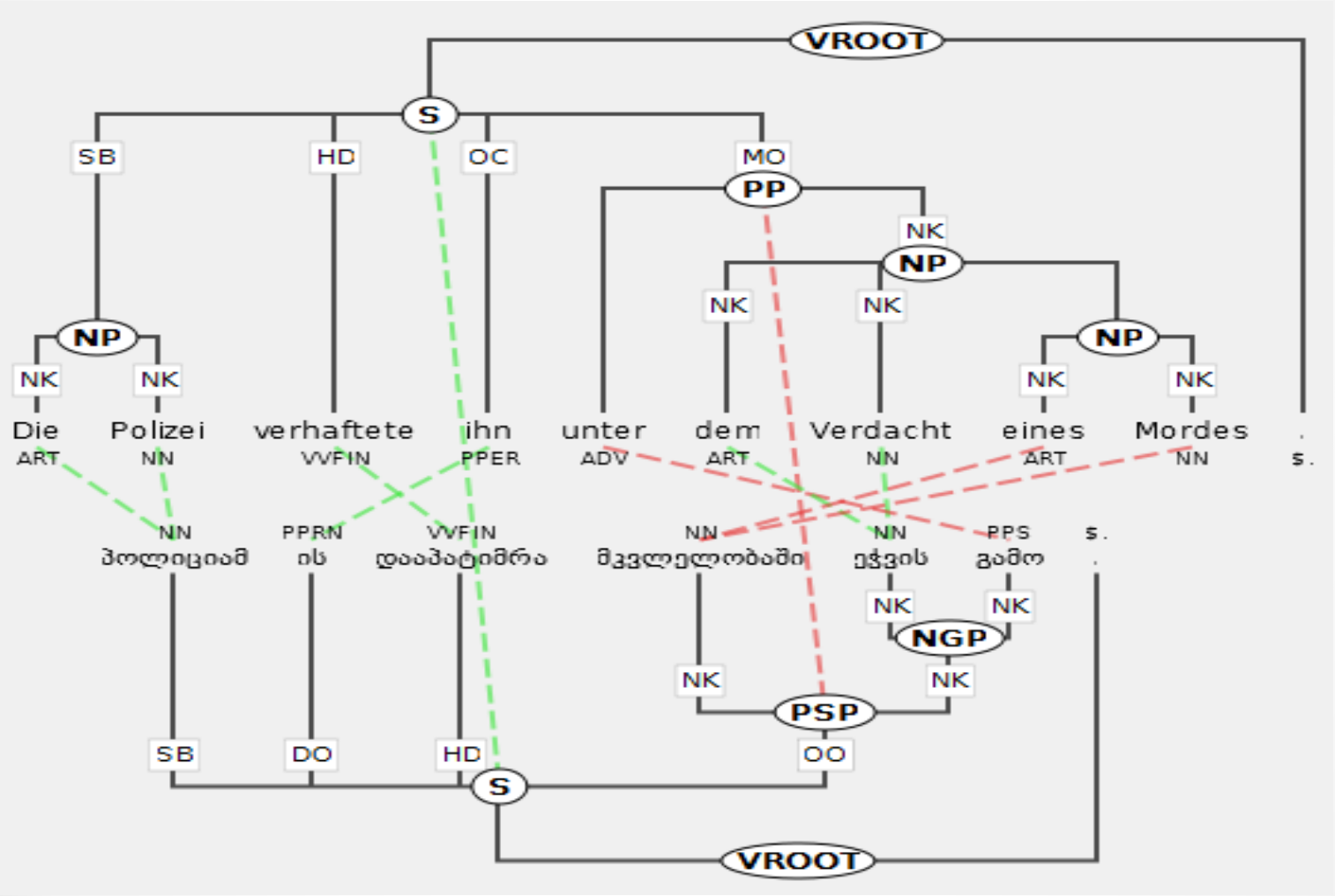
. Phrases shall be aligned only if the tokens, that they span, represent the same meaning and if they could serve as translation units outside the current sentence context

.The grammatical forms of the phrases need not fit in other contexts, but the meaning has to fit

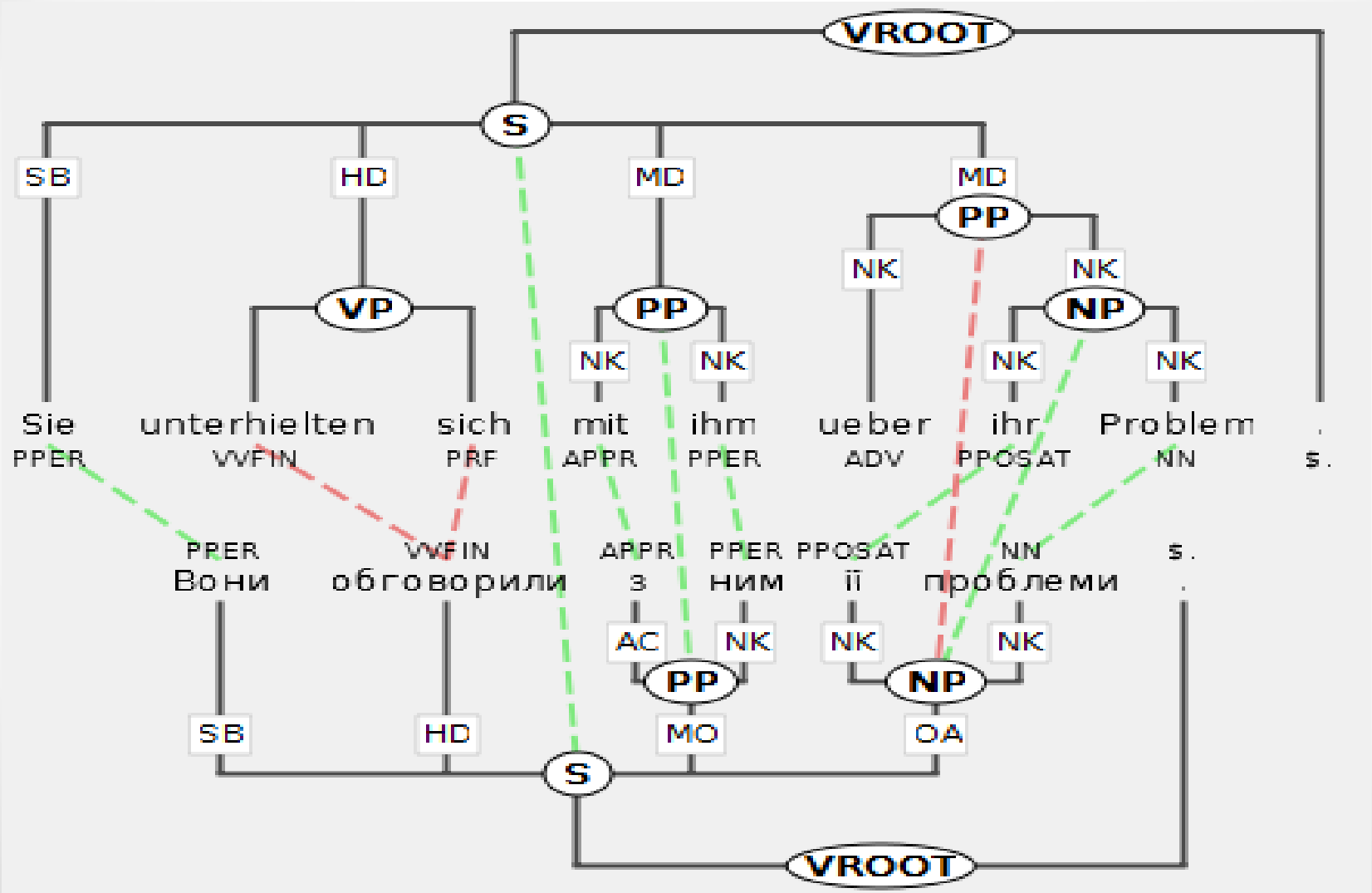
.The Stockholm TreeAligner guidelines allows phrase alignments within m : n sentence alignments and 1 : n phrase alignments

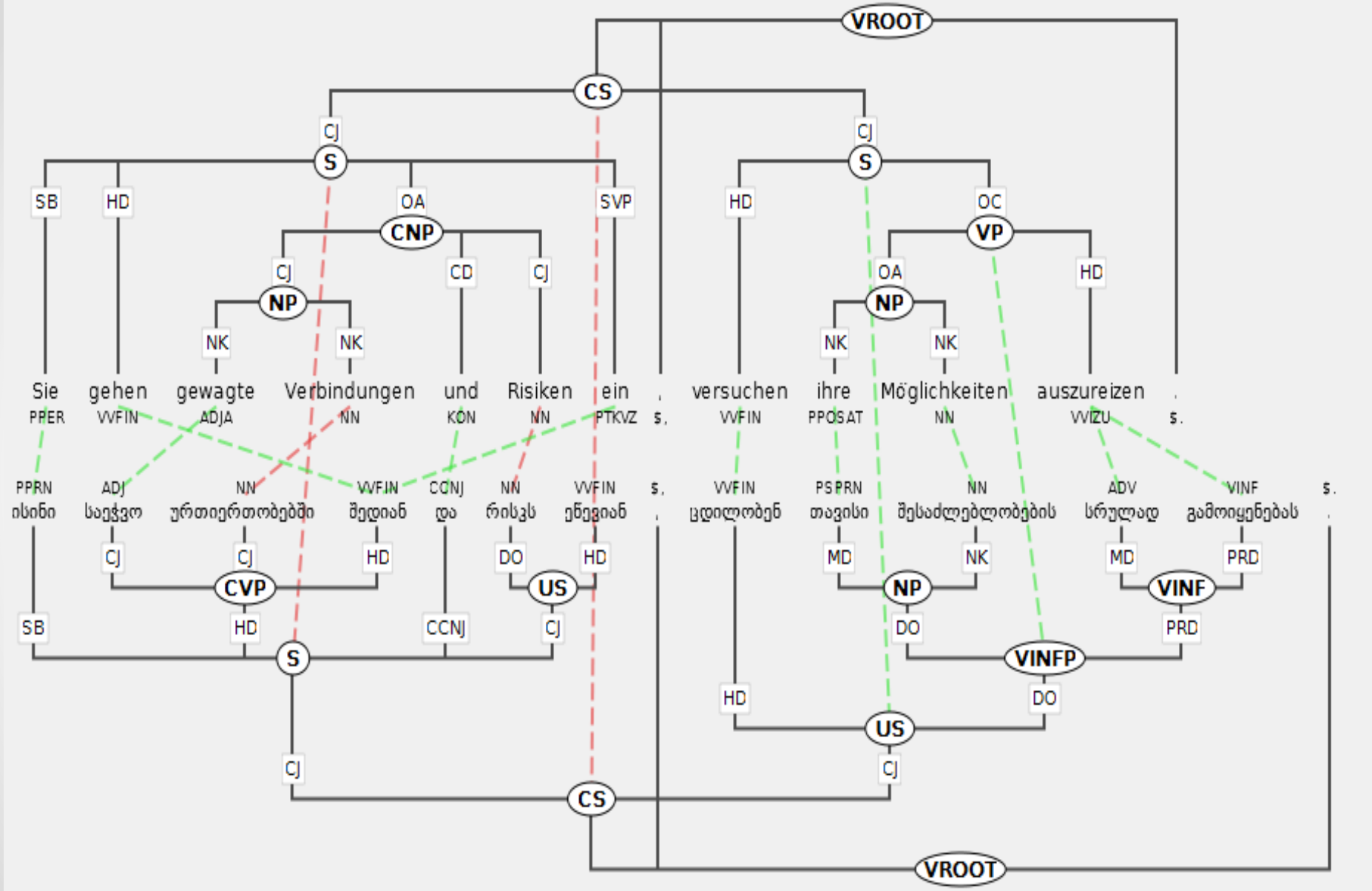
. If nodes and words represent just approximately the same meaning, they are aligned as fuzzy translation correspondences by means of lines in the red colour

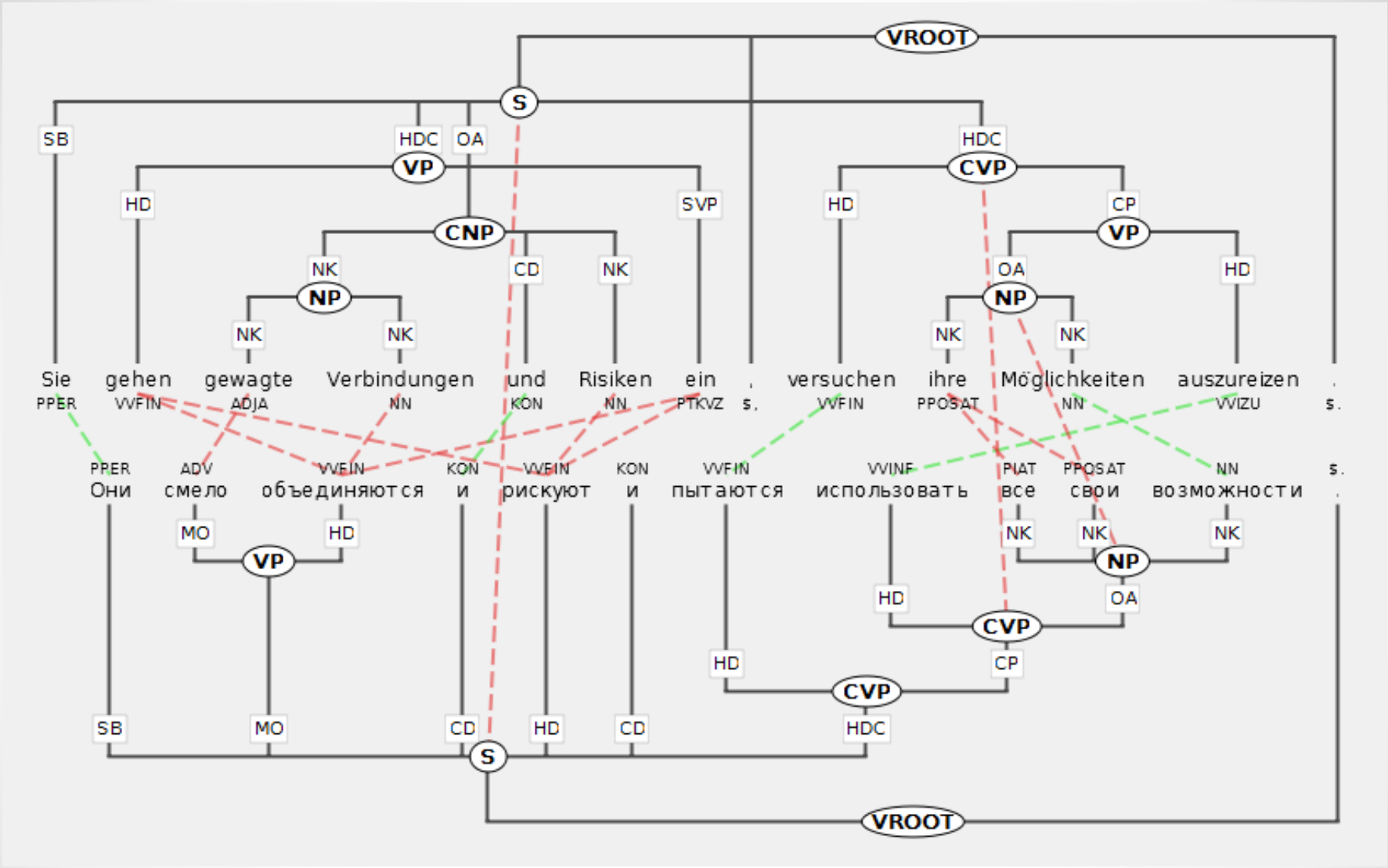
“Die Polizei verhaftete ihn unter dem Verdacht eines Mordes”
(lit. The police arrested him under a suspicion of a murder)

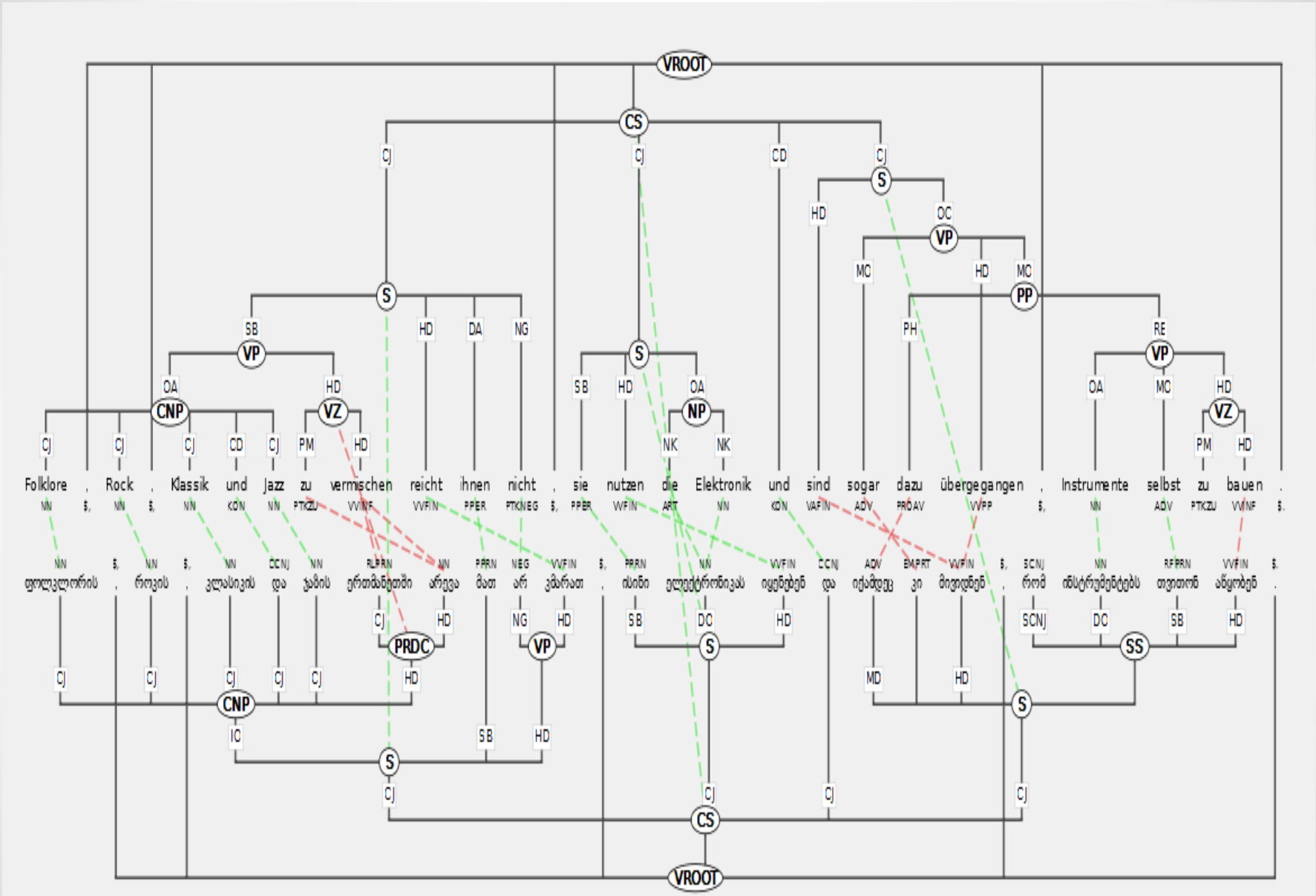


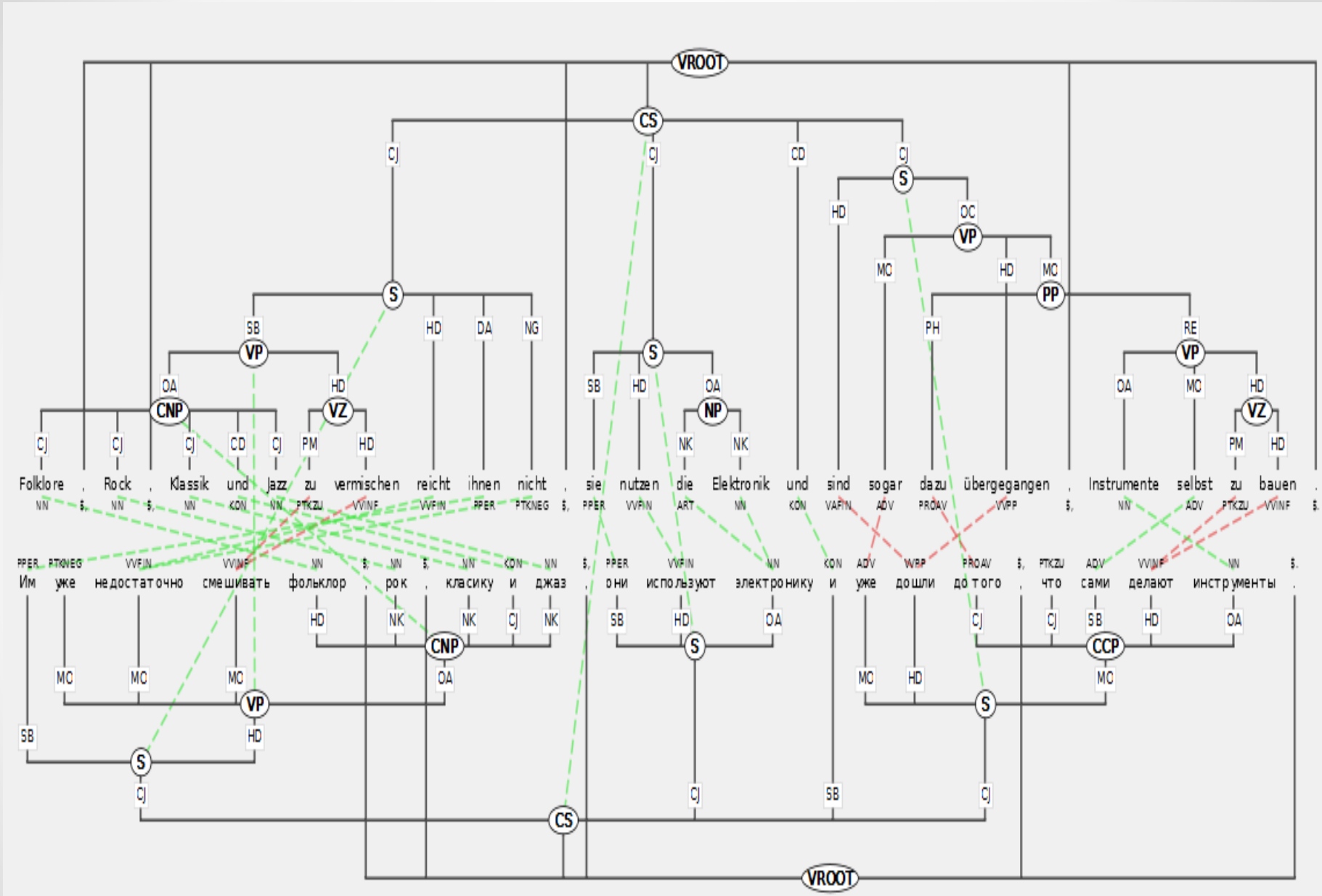
“Sie unterhielten sich mit ihm über ihr Problem”
(lit. They discussed with him (about) her problem)

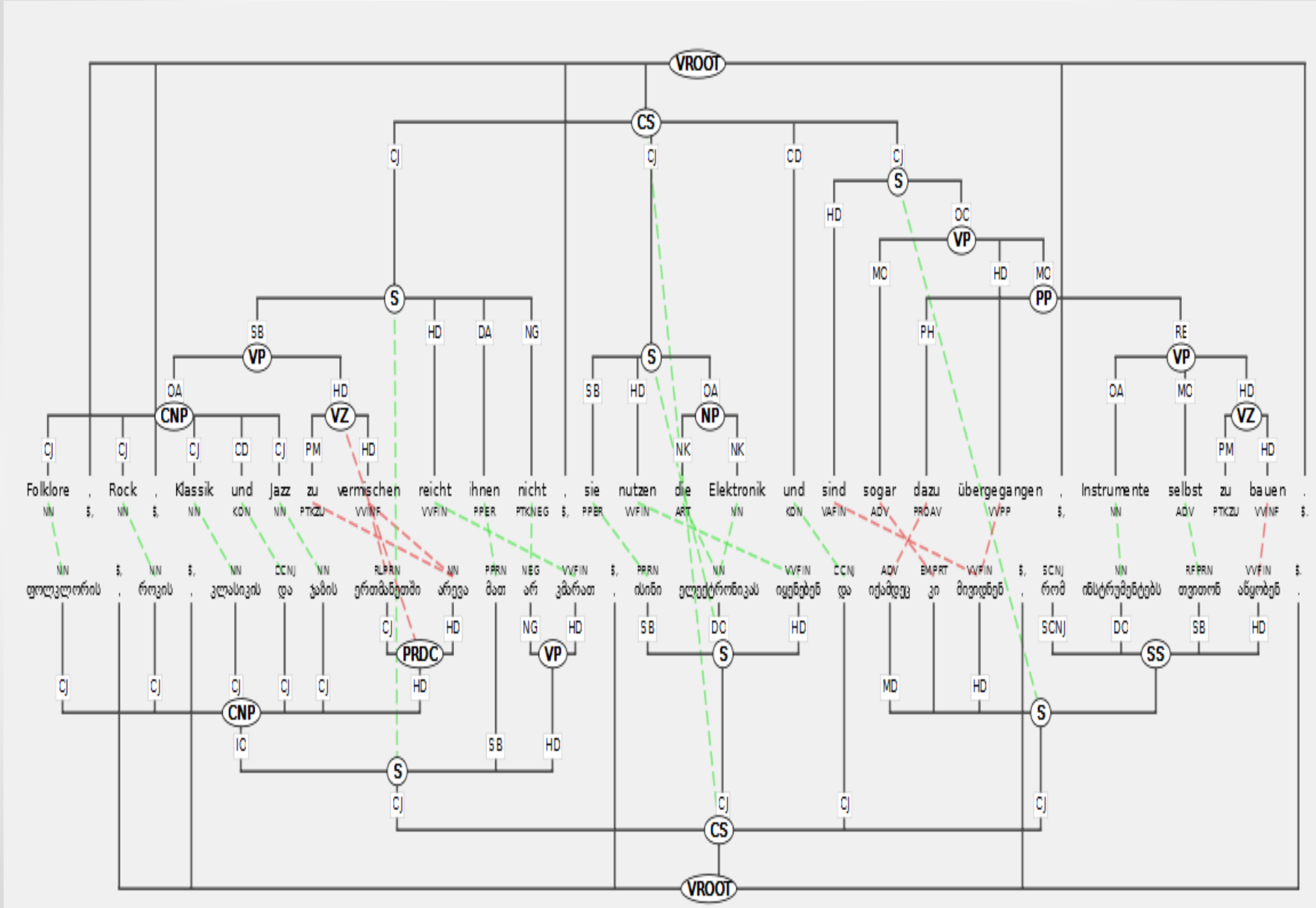



VROOT
CS
CJ
S
SB
HD
OA
SVP
CNP
CJ
CD
CJ
NP
NK
NK
Sie
gehen
gewagte
Verbindungen
und
Risiken
ein
,
PPER
VVFIN
ADJA
NN
KON
NN
PTKVZ
$,
PPRN
ისინი
ADJ
საეჭვო
NN
ურთიერთობებში
VVFIN
შედიან
CCNJ
და
NN
რისკს
VVFIN
ეწევიან
$,
CJ
CJ
HD
DO
HD
CVP
US
SB
HD
CCNJ
CJ
S
CJ
CS
CJ
S
HD
OC
VP
OA
HD
NP
NK
NK
versuchen
ihre
Möglichkeiten
auszureizen
.
VVFIN
PPOSAT
NN
VVIZU
$.
VVFIN
ცდილობენ
PSPRN
თავისი
NN
შესაძლებლობების
ADV
სრულად
VINF
გამოიყენებას
$.
HD
MD
NK
MD
PRD
NP
VINF
DO
PRD
VINFP
HD
DO
US
CJ
VROOT


VROOT
S
SB
HDC
OA
VP
HD
CNP
SVP
NK
NP
CD
NK
NK
NK
Sie
PPER
gehen
VVFIN
gewagte
ADJA
Verbindungen
NN
und
KON
Risiken
NN
ein
PTKVZ
,
$,
versuchen
VVFIN
HDC
CVP
HD
CP
VP
OA
NP
HD
NK
NK
ihre
PPOSAT
Möglichkeiten
NN
auszureizen
VVIZU
.
$.
PPER
Они
ADV
смело
VVFIN
объединяются
KON
и
VVFIN
рискуют
KON
и
VVFIN
пытаются
VVINF
использовать
PIAT
все
PPOSAT
свои
NN
возможности
$.
MO
HD
VP
NK
NK
NK
NP
HD
OA
CVP
HD
CP
CVP
SB
MO
CD
HD
CD
HDC
S
VROOT


VROOT
CS
CJ
CJ
CD
CJ
S
HD
OC
VP
MO
HD
MO
PP
PH
RE
VP
OA
MO
HD
VZ
PM
HD
S
SB
HD
DA
NG
VP
OA
HD
CNP
VZ
S
SB
HD
OA
NP
NK
NK
CJ
CJ
CJ
CD
CJ
PM
HD
Folklore , Rock , Klassik und Jazz zu vermischen reicht ihnen nicht , sie nutzen die Elektronik und sind sogar dazu übergegangen , Instrumente selbst zu bauen .
NN $, NN $, NN KON NN PTKZU VVINF VVFIN PPER PTKNEG $, PPER VVFIN ART NN KON VAFIN ADV PROAV VVPP $, NN ADV PTKZU VVINF $.
ფოლკლორის , როკის , კლასიკის და ჯაზის ერთმანეთში არევა მათ არ ეკმარათ , ისინი ელექტრონიკას იყენებენ და იქამდეც კი მივიდნენ , რომ ინსტრუმენტებს თვითონ აწყობენ .
NN $, NN $, NN CCNJ NN RLPRN NN PPRN NEG VVFIN $, PPER NN VVFIN CCNJ ADV EMPRT VVFIN $, SCNJ NN RFPRN VVFIN $.
CJ HD
PRDC
NG HD
VP
SB DO HD
S
ADV EMPRT VVFIN
SCNJ DO SB HD
SS
CJ CJ CJ CJ CJ HD
MD HD
CNP
S
IC
SB HD
S
CJ CJ CJ CJ
CS
VROOT


VROOT
CS
CJ
CJ
CD
CJ
S
HD
OC
VP
MO
HD
MO
PP
PH
RE
VP
OA
MO
HD
VZ
PM
HD
S
SB
HD
DA
NG
VP
OA
HD
CNP
VZ
S
SB
HD
OA
NP
NK
NK
CJ
CJ
CJ
CD
CJ
PM
HD
Folklore , Rock , Klassik und Jazz zu vermischen reicht ihnen nicht , sie nutzen die Elektronik und sind sogar dazu übergegangen , Instrumente selbst zu bauen .
NN $, NN $, NN KON NN PTKZU VVINF VVFIN PPER PTKNEG $, PPER VVFIN ART NN KON VAFIN ADV PROAV VVPP $, NN ADV PTKZU VVINF $.
PPER PTKNEG VVFIN VVINF NN $, NN $, NN KON NN $, PPER VVFIN NN KON ADV VVPP PROAV $, PTKZU ADV VVINF NN $.
Им уже недостаточно смешивать фольклор , рок , класику и джаз , они используют электронику и уже дошли до того , что сами делают инструменты .
HD NK NK CJ NK SB HD OA CJ CJ SB HD OA
CNP S CCP
MO MO MO OA MO HD MO
VP S
SB HD
S
CJ CJ SB CJ
CS
VROOT


VROOT
CS
CJ
CJ
CD
CJ
S
HD
OC
VP
MO
HD
MO
PP
PH
RE
S
SB
HD
DA
NG
S
SB
HD
OA
NP
NK
NK
VP
OA
HD
CNP
VZ
OA
MO
HD
VZ
CJ
CJ
CJ
CD
CJ
PM
HD
PM
HD
Folklore , Rock , Klassik und Jazz zu vermischen reicht ihnen nicht , sie nutzen die Elektronik und sind sogar dazu übergegangen , Instrumente selbst zu bauen .
NN $, NN $, NN KON NN PTKZU VVINF VVFIN PPER PTKNEG $, PPER VVFIN ART NN KON VAFIN ADV PROAV VVPP $, NN ADV PTKZU VVINF $.
NN $, NN $, NN CCNJ NN RLPRN NN PPRN NEG VVFIN $, PPRN NN VVFIN CCNJ ADV EMPRT VVFIN $, SCNJ NN RFPRN VVFIN $.
ფოლკლორის , როკის , კლასიკის და ჯაზის ერთმანეთში არევა მათ არ კმარათ , ისინი ელექტრონიკას იყენებენ და იქამდეც კი მივიდნენ , რომ ინსტრუმენტებს თვითონ აწყობენ .
CJ
HD
NG
HD
SB
DC
HD
SCNJ
DC
SB
HD
PRDC
VP
S
SS
CJ
CJ
CJ
CJ
CJ
HD
MD
HD
CNP
S
IC
SB
HD
S
CJ
CJ
CJ
CJ
CS
VROOT

Divergences between the GRUG languages

. **GE, RU, UA** – inflectional languages
. **GO** - an aglutinative language

. **RU, UA, GO -** absence of articles

. **Word Order Freedom:**

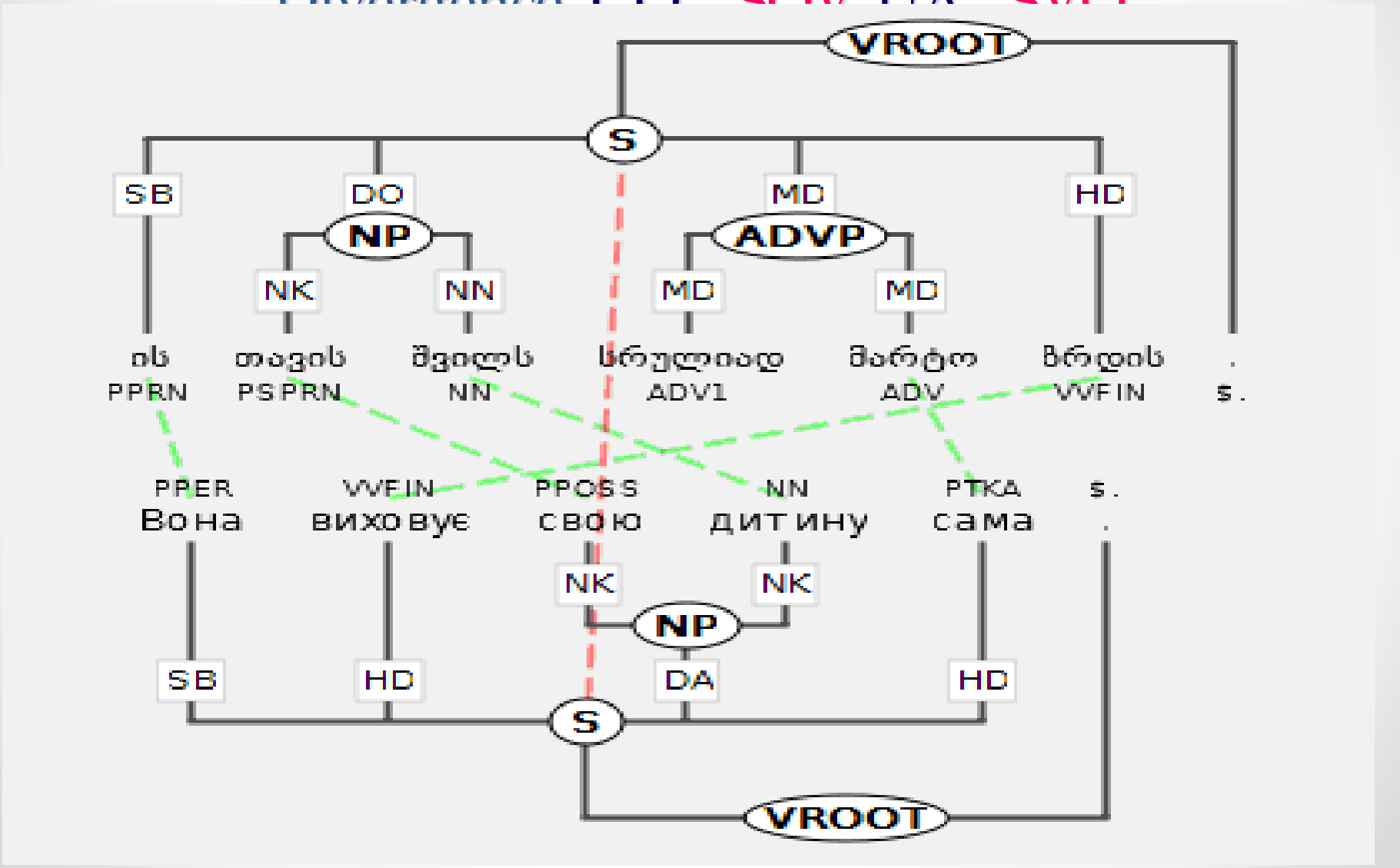
GE - **SOV** in dependent clause and **SVO** in main clause;

GO - has a free word order as a result of its rich morphological structure but a preferred basic word order without a Theme/Rheme bias for Georgian is **SOV**, which is canonical for the German dependent clauses

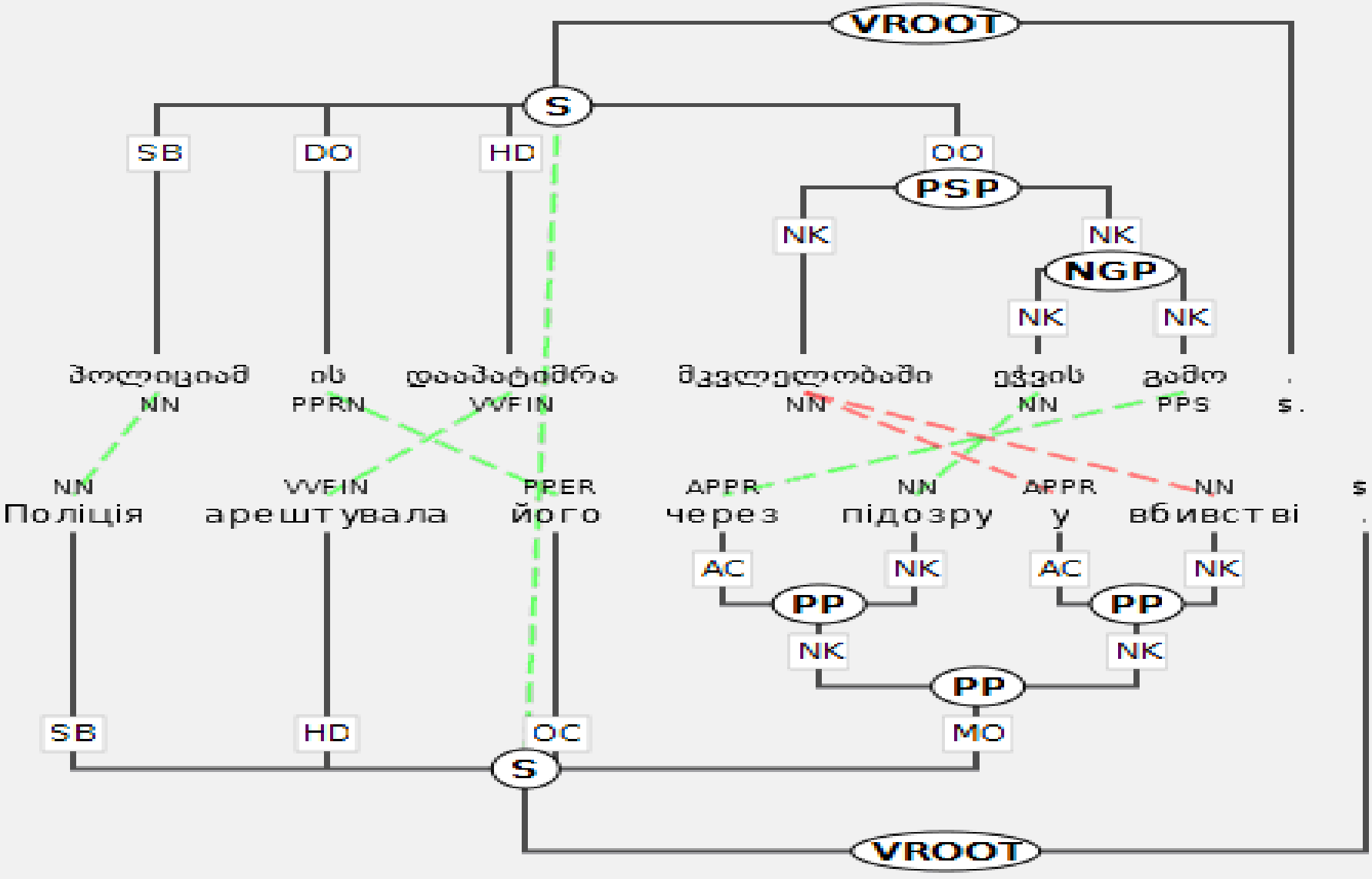
RU, UA - have a relative free word order, though, to a lesser extent than it is observed in the Georgian language and a preferable word order is **SVO** which is canonical for the German main clauses

Despite the mentioned difference, a 1:1 alignment on word, phrase and sentence level can be often viewed in the GRUG parallel trees

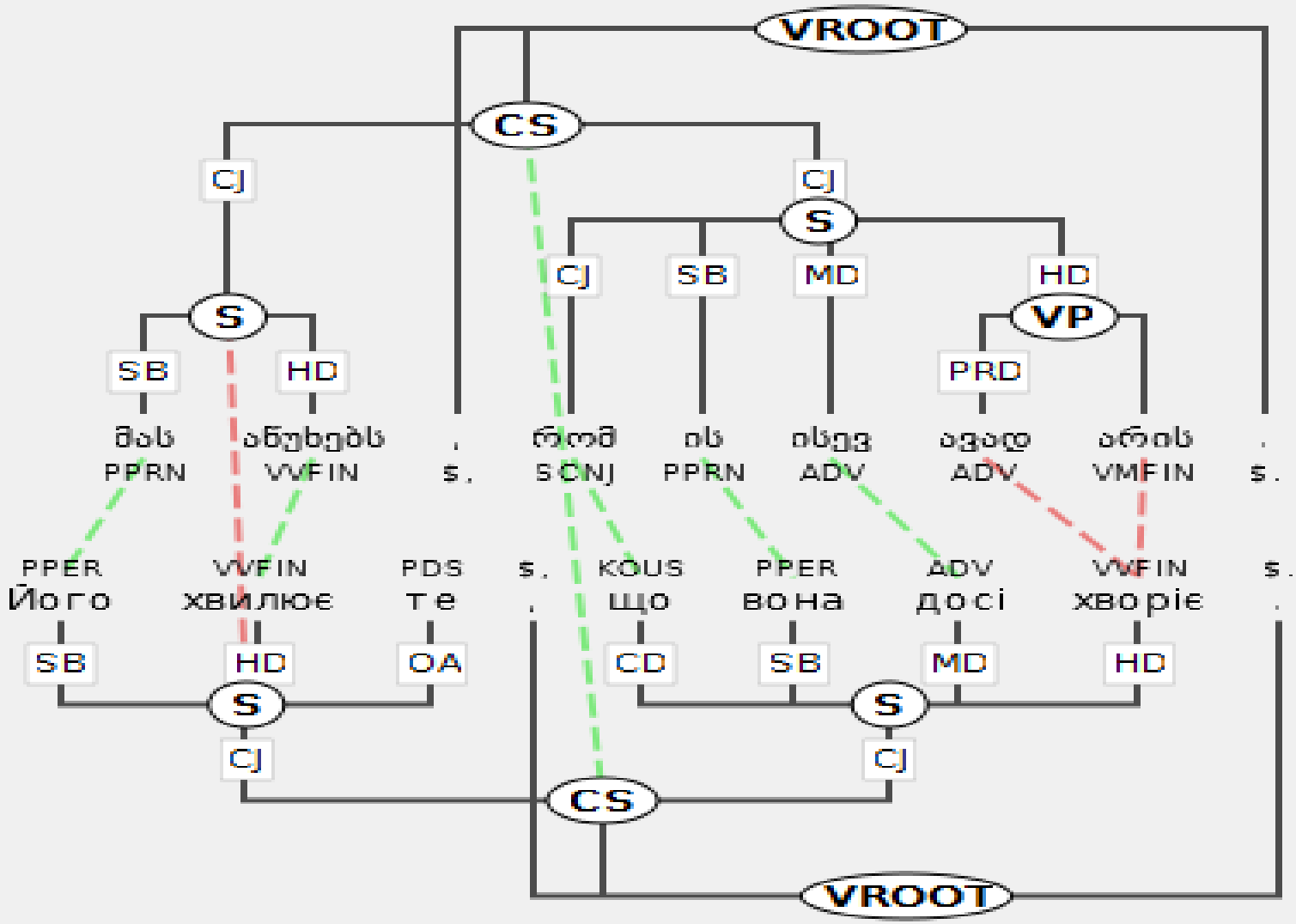
Divergence GO – SOV UA – SVO



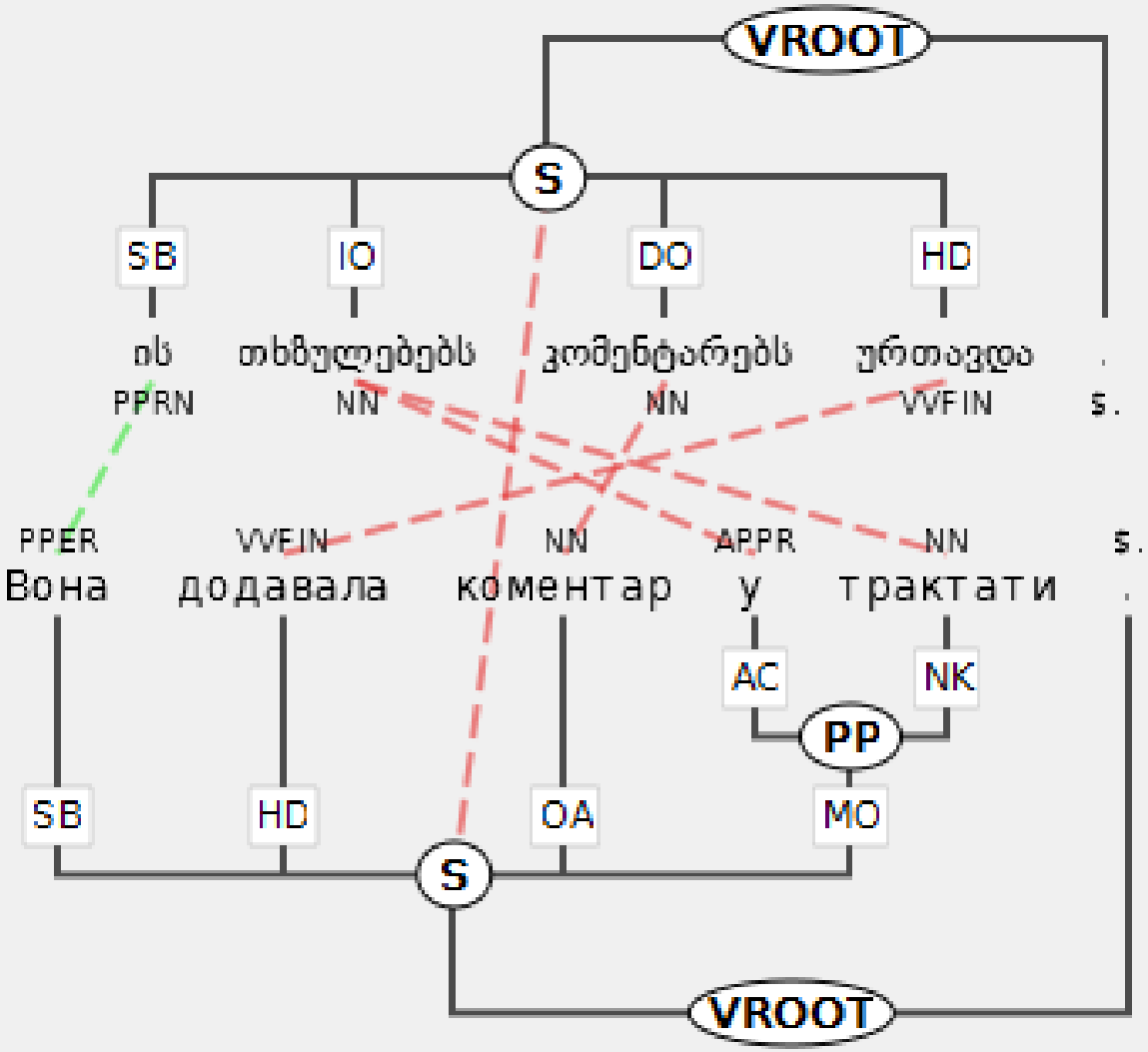
Divergence on a Phrase level



Divergence on a predication level



Divergence on the level of Pragmatics



Conclusions

- **Despite the outlined typological differences between the GO, RU, UK, GE languages, the alignment tools for monolingual (*Synpathy*) and bilingual TreeBanks (*the Stockholm TreeAligner*) are capable to cope with divergences in linguistic structures and visualize structural and translation equivalents in GRUG parallel corpora**

The developed mono and bilingual TreeBanks are useful in

- **Translation Studies:** for Visualization of Contrastive Phenomena Between a Pair of Languages
- **Corpus Linguistics**: as Resources for Research and Analysis for Syntactic Structures of a Pair of Languages
- **Language Engineering**: for Testing and Evaluation of Automatic Parsers
- **Integrated Technologies for Translation**: as Training Material for an Example-Based Machine Translation and as a Database for Translation Memory Systems

References

Brants, Sabine and Hansen, Silvia (2002). Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002),* pp. 1643–1649, Las Palmas.

Kapanadze, Oleg, Kapanadze, Nunu, Wanner, Leo and Klatt, Steffan (2002). Towards A Semantically Motivated Organization of A Valency Lexicon for Natural Language Processing: A GREG Proposal. In *Proceedings of the EURALEX conference,* Copenhagen.

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

Kapanadze, Oleg (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, Oleg (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.

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