International Journal of Computational Linguistics and Applications
vol. 7, no. 1, 2016, pp. 85–100
Received 09/01/2015, accepted 21/08/2015, final 09/02/2016
ISSN 0976-0962, http://ijcla.bahripublications.com

Parallel TreeBanks: Observations for Implication of Equivalent Alignments

OLEG KAPANADZE

Tbilisi State University, Tbilisi, Georgia

ABSTRACT

Building a parallel Treebank anticipates alignment of linguistic information represented by diverse structures on different layers of a bilingual text. In this paper, we describe our observations for inference translation equivalents in parallel texts of languages with diverse structures - German and Georgian. They belong to the different language families and as a consequence enjoy different typological features manifested by diverse morphological structures, word and phrase order in a clause. In the bilingual German-Georgian Treebank development process it has been given a try to cluster the tolerant syntactic structures and classify phrase conventional translations that could be considered as equivalent units in the bilingual text alignment issue.

KEYWORDS: *under-resourced languages, Treebanks, annotation, alignment.*

1 Introduction

In this paper we outline a study carried out in the framework of a Multilingual GRUG project having been intended for building a German-Georgian, a German-Russian, a German-Ukrainian and a Georgian-Ukrainian parallel treebanks. The languages (except German) involved in the project are under-resourced languages [4] (cl. http://fedora.clarin-d.uni-saarland.de/grug/).

A significant part of modern treebanking literature is devoted to creation of large Treebanks for the languages with a relatively simple morphology and relatively fixed word order. Data-driven treebanking is now at the state where naturally occurring text in the news domain can be automatically annotated with high accuracy according to standard parsing evaluation measures. However, when moving from languages with relatively fixed word order to languages with richer morphologies and less-rigid word orders, the standard issues for annotation of bilingual Treebanks developed for languages with fixed word order exhibit a large drop in accuracy.

To overcome this obstacle, in the initial phase the GRUG project had been concentrated on development of a parallel treebank for a typologically dissimilar language pair - German and Georgian [5] (cl. http://clarino.uib.no/iness). The later is an agglutinative language using in wordform building both, suffixing and prefixing. In German, word-order is relatively fixed, while in Georgian as in many other languages, word order is much more flexible (for example, the subject may appear either before or after a verb, etc.). In languages with flexible word order, the meaning of the sentence is realized using other structural elements, like word inflections or markers, which are referred to as morphological information.

Morphology provides useful hints for resolving syntactic ambiguity, and the parsing model should have a way of utilizing these hints. The Georgian text morphological annotation, tagging and lemmatizing procedures were done with a finite-state morphological transducer based on the XEROX FST tools [6], [7]. A lexicon-based parse engine has been oriented to capture the specifics of the Georgian morphology manifesting rich syntactic clues encapsulated in the finite verb forms. The monolingual Treebank syntactic annotation adhered the German TIGER project guidelines [1]. The tagset for Georgian follow the STTS/Stuttgart-Tübingen-Tagset scheme with necessary changes relevant for the Georgian grammar formal description.

The POS-tagged and lemmatized parallel German-Georgian texts were syntactically annotated manually by means of *the Synpathy* software [14]. An output of the monolingual syntactic annotation issue is in the TIGER-XML format.

The alignment of the monolingual syntactically annotated trees into parallel Treebanks had been accomplished using *the Stockholm TreeAligner*, a tool for work with parallel treebanks which inserts

alignments between the mirrored pairs of syntax trees [12], [13]. *The Stockholm TreeAligner* uses monolingual graph structures in the TIGER-XML format as representations and handles in parallel treebanks alignment of tree structures in addition to the token alignment.

A study of the similar projects for building the parallel treebanks [3], [9], [10], [11], evinced that an alternative approach has been also advocated for some agglutinative languages. In a Quechua-Spanish parallel Treebank, due to highly inflectional structure, the monolingual Quechua treebank had been annotated on morphemes rather than words and a Role and Reference Grammar has been opted for its annotation. This allowed to link morpho-syntactic information precisely to its source.

Despite the significant importance of morphology for the Georgian syntactic parsing, we believe, there is no need to annotate the Georgian Treebank on morphemes to capture its syntactic peculiarities. In the following chapters we intend to outline some of those features necessary for equivalent alignment implications.

2 Structural Divergences and Some Alignment Inferences

A notable typological difference between the German and the Georgian languages is absence of articles as grammatical category in the later. Its general functions in Georgian take over as certain lexical items (Pronouns), as well as some grammatical means.

From the structural view point, a significant syntactic divergence already have been sketched, is the word order freedom in two languages. For the German language there is an assumed basic word order, which is postulated to be either SOV in dependent clauses and SVO in main clauses. Quite frequently, within those statements, predictions about the Subject have been replaced by predictions about a general pre-verbal position, yielding XOV/XVO for German.

On a contrary, in Georgian the linguists admit a relative free word order as a result of its rich and productive morphology. Nevertheless, a preferred basic word order without a Theme/Rheme bias for Georgian is SOV, which is canonical for the German dependent clause.

The most notable divergence in a syntactic description model for the Georgian clause is a phenomenon classified as a mutual

government and agreement relations between verb-predicate and its actants, which number may reach up to three in a single clause. This anticipates control of the noun case markers by verbs, whereas the verbs in their turn, are governed by nouns with respect to a grammatical person and number. As a consequence of the verb-predicate capability to reflect morphologically the agreement relations, pronouns as actants (Subject (Sb), Direct Object (DO), Indirect Object (IO)) can be omitted in the word order without a consequence for the clause meaning comprehension. The "reduced" clauses are equally "sufficient/eligible" as their source ones in terms of the clause meaning representation.

In Figure 1 is depicted a complex sentence (CS)

თუ ღმერთი გწამთ, არ მითხრათ ახლა, რომ შავი თეთრია

(lit. If you believe in god (=For god's sake), do not tell me now that black is white).

It has two clauses CnS (a conditional clause) and MS (a main clause) that are the reduced ones of their source variants.

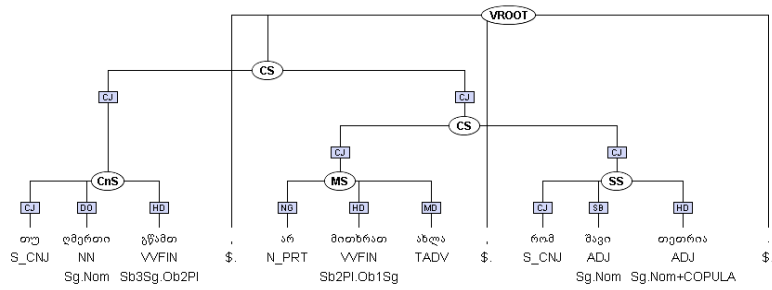

Fig. 1. A syntactically annotated Georgian sentence in the TIGER-XML format.

The sentence in Figure 1 visualizes a hybrid approach to the syntactic annotation procedure as tree-like graph structures and integrates annotation according to the constituency representations and functional relations. Consequently, in a tree structure the node labels are phrasal categories, whereas the parental and secondary edge labels correspond to syntactic functions.

A non reduced option of the CnS and the MS with the respective valency relations would yield to the following syntactic sub-graphs:

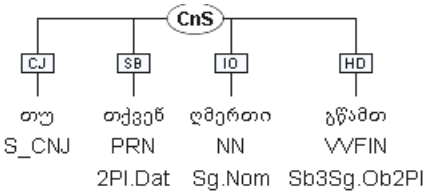

Fig. 2. A conditional clause (CnS) sub-graph of the syntactically annotated sentence from Figure 1.

The head (HD) or the kernel of the CnS გწამთ ("you believe it") is POS-tagged on the morphological level as a finite verb form (VVFIN) manifested as the Subject 3rd person singular (Sb3Sg) and Object 2nd person plural (Ob2Pl), though, in a consequent reduced clause PRN (Ob2Pl) has been dropped (cl. CnS in Figure 1).

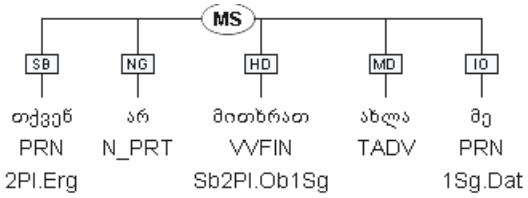

Fig. 3. A conditional clause (MS) sub-graph of the syntactically annotated sentence from Figure 1.

The head (HD) of the MS node მითხრათ ("you will/would tell me") is POS-tagged as VVFIN of Subject 2rd person plural (Sb2Pl) and Object 1st person singular (Ob1Sg). The consequent reduced clause in Figure 1 lacks Subject (SB) as PRN 2nd person plural in Ergative case (2Pl.Erg) and Indirect Object (IO) as PRN 1st person singular Dative case (1Sg.Dat).

A syntactically annotated German translation equivalent of the Georgian sentence from Figure 1 is visualized in Figure 4.

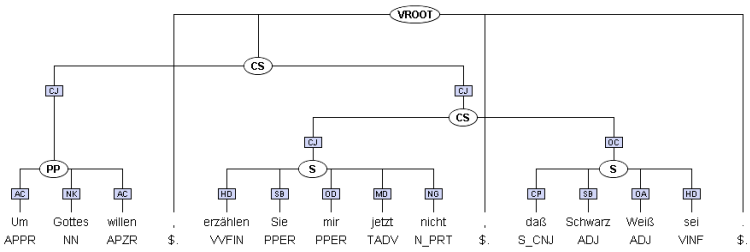

Fig. 4. A syntactically annotated German translation equivalent of the Georgian sentence from Figure 1

The monolingual Georgian and German syntactically annotated clauses as tree-like graphs are fed to *the Stockholm TreeAligner* engine. An output of mirrored parallel monolingual trees are visualized by the graphical viewer. After manual alignment of tokens and phrases the resulted parallel trees are depicted in Figure 5.

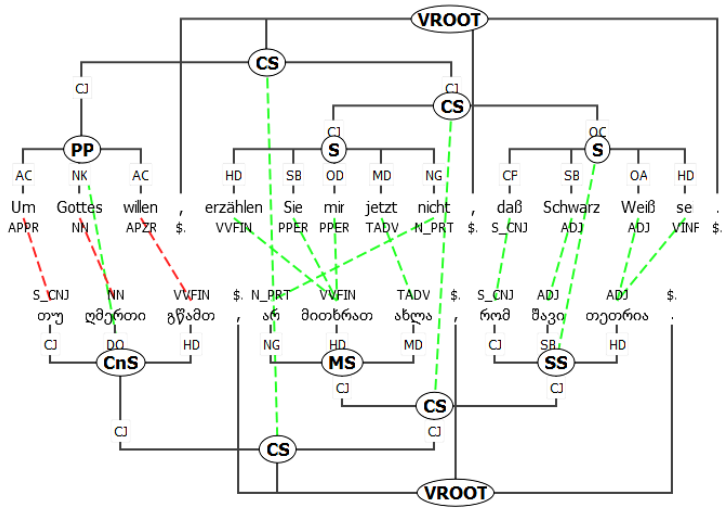

Fig. 5. A German-Georgian syntactically annotated and aligned tree.

The nodes and words from two languages with the same meaning are aligned as exact alignments using the green color. If nodes and words from one language represent just approximately the same

meaning in the other language, they are aligned as "fuzzy" translation equivalents spanned by lines in the red color [12], [13].

Phrase alignment, as an additional layer of information on top of the syntax structure, shows which part of a sentence in one language is equivalent to a part of a parallel sentence in the other language. This is done with help of a graphical user interface of *the Stockholm TreeAligner*. The phrases are aligned only if the tokens, that they span could serve as translation units outside the current sentence context. *The Stockholm TreeAligner* guidelines allow phrase alignments within m : n sentence alignments and 1 : n phrase alignments. The grammatical forms of the phrases need not fit in other contexts, but the meaning has to fit.

Due to divergence between the German clause and its reduced Georgian counterpart we witness n:1 token alignment which theoretically is a "fuzzy" alignment case and the tokens must be spanned by lines in the red colour. Nevertheless, we believe this is a good alignment, hence, marked in the consequent colour.

Another case to be discussed is the first clause which is classified as PP (Prepositional Phrase) in German. The corresponding Georgian one is a clause with a verb-predicate as its head. The tokens are linked with the red lines, since they can not be counted as translation equivalents outside of the current context. Nevertheless, these two structural units are accepted as conventional translation equivalents on the pragmatic level, and therefore are recognized as a "good" alignment. In general, we adhere a basic principle that all possible idiomatic phrases with their counterparts in parallel text should be classified as good alignment pairs.

3 Prepositional vs Postpositional Phrases

An implication of constituent and word order typological dissimilarity between the German and the Georgian languages can be observed in respective syntactic structures. One of the points discussed further concerns German prepositional phrases (PP) which are headed by prepositions standing in different places.

3.1 Prepositional Phrases in German

A lexical class of prepositions in the German Language comprises as

morphologically simple *(in, an, auf, zu, für, mit),* as well as the complex units *(aufgrund, anstelle, unweit, in Bezug auf).* Their aggregated number reaches a hundred entities [2]. But, if we count also the verb participle forms, such as *während, entsprechend, betreffend, ausgenommen, ungeachtet, unbeschadet,* their number will increase significantly.

Every prepositional phrase contains a preposition as "a lexical head" or a headword linked with complements which in their turn may represent different phrasal categories: Noun Phrase (NP), Adjective Phrase (AP), Adverbial Phrase (AdvP). It might be also a Prepositional Phrase (PP) as a complement to the head preposition.

Like Noun Phrases the subtrees of Prepositional Phrases are described as a node with "horizontal" constituents having been displayed on the same level. There are three options for placing prepositions with regard to the complement in the Prepositional Phrase:

Front position: E.g. preposition *für* ("for") tagged as APPR:

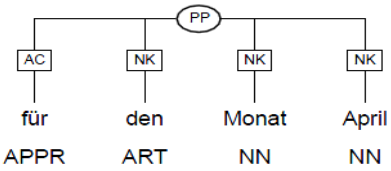

Fig. 6. A preposition preceding a Noun Phrase in a PP (lit. "for the month April").

Post position: E.g. preposition *nach* ("according", "after") tagged as APPO:

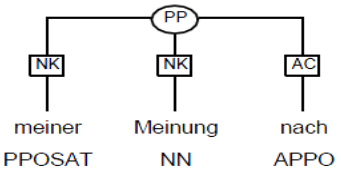

Fig. 7. A preposition following a Noun Phrase in PP (lit. "[to] my opinion according").

Circumposition: E.g. *um…willen* ("for … sake") tagged as APPR…APZR

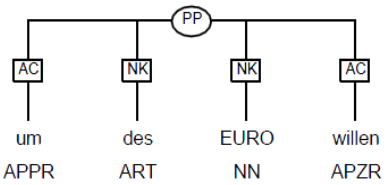

Fig. 8. A preposition in a circumposition to a Noun Phrase in PP (lit. "for the EURO sake").

Pursuant to the TIGER annotation tagset in each three cases for the preposition an edge label is AC (Adpositional Case marker).

PPs on the sentence level in most cases are Modifiers (MD) [15] as in the following examples:

um 5 Uhr ("at 5 o'clock")

wegen der Ankunft der Delegation ("due to the advent of the delegation")

nach London ("to London")

in diesem Durcheinander ("in this mess").

For all of them a subtree has the same flat structure as it has been for the last prepositional phrase:

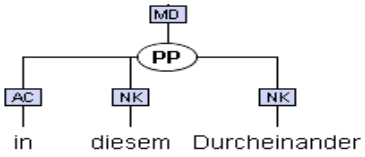

Fig. 9. A typical subtree of Prepositional Phrase as syntactic Modifier (MD).

Prepositional Phrases can also be assigned the syntactic function of the Prepositional Object (OP) as it can be observed in the example below:

Ich entschuldige mich vielmals für die Störung

(lit. "I apologize [many times] for the disruption")

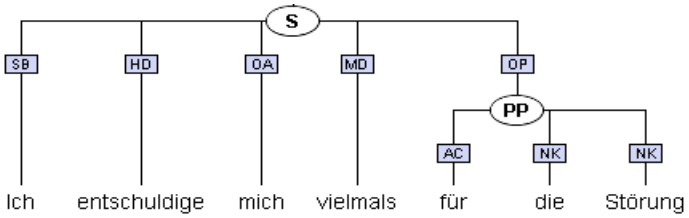

Fig. 10. A Prepositional Phrase as Prepositional Object (OP).

Prepositional phrases as Prepositional Objects enjoy the same tree structure as prepositional phrases with the function of a syntactic modifier. According to Klenk [8], they could be distinguished by exploring a "behavior" of the preposition in the prepositional phrase. In the case of the Prepositional Object the prepositions in the respective phrase loose their lexical meaning and acquire some functional role instead. Therefore, the verb in the phrase determines a specific preposition that can conform to the phrase.

3.2 Postpositional Phrases in Georgian Language

In Georgian the translation equivalents for the German Prepositional Phrases are Postpositional Phrases (PSP). In PSP some postpositions, as independent unchangeable words, stand on their own and appear after noun. Some others adhere to the noun base form as an enclitic particle. Nevertheless, a German PP in Georgian can be also translated by a phrase headed by a noun with a case inflection as it can be observed in a German sentence from Figure 11 and the alligned Georgian counterpart:

Er verwöhnt sie mit Blumen (lit. "He cossets her with flowers")

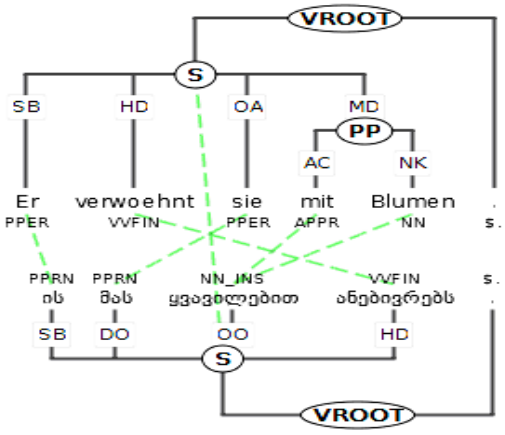

Fig. 11. Divergence on a phrasal/constituent structure alignment level for German and Georgian.

Besides the divergence in syntactic category labels, these constituents also differ from functional view point. In the German grammar they are considered as Modifiers (MD), whereas in Georgian the PSPs traditionally are classified as "ordinal objects" (OO). They differ from direct (DO) and indirect objects (IO) also formally, since later two are marked morphologically by the specific affixes in verb, which is not the case with OO.

The dicussed structural difference can be disregarded in the alignment process and the German PP "*mit Blumen*" (lit. "with/by means of Flower") considered as a "good" translation equivalent, though, a 2:1 alignment on a token level. The suggested solution derives from a prerequisit of "translation equivalence outside the current sentence context". In the other words, a German PP

mit +N (*mit Blumen*)

is always translated in Georgian as:

N+Instrumental_case (ყვავილებ_ით)

However, there are options when the German PPs are aligned to the Georgian PSP as "good" alignments on the phrase level, though, again with a 2:1 alignment on the token leyer as in the Figure 12 for the German sentence

Sie unterhielten sich mit ihm über ihr Problem

(lit. “They discussed with him [about] her problem“),
where the German PP

[*über Problem*]

is aligned to Georgian PSP

[NN+postp ~ პრობლემა+ზე]:

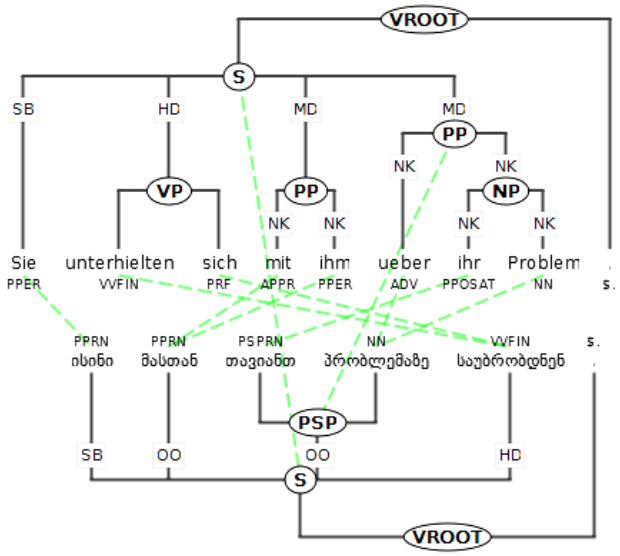

Fig. 12. A “good” alignment on a constituent structure/phrasal level for German and Georgian.

Nevertheless, we count them to be a “good” alignment, since they can serve as translation equivalents outside of the current sentence context.

In contrary, in Figure 13 is presented an example of a “good” alignment between the German and the Georgian counterpart for a sentence

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

though, almost all of the tokens are spanned as “fuzzy” translation unites in the parallel syntactically annotated trees.

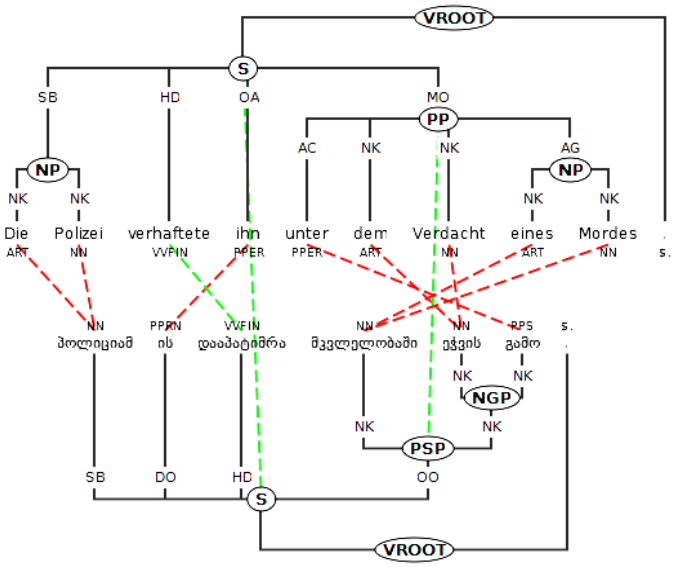

Fig. 13. An example of a German PP and a Georgian PSP with "good" alignment.

The last five tokens from the German sentence, are spanned as "fuzzy" alignments to the three Georgian counterparts, though, on the phrasal level they are accepted as "good" alignment of PP to PSP. A preposition *unter* ("under") and a NP e*ines Mordes* ("a murder") from the German PP are aligned to the Georgian counterparts გამო and მკვლელობაში from a PSP as "fuzzy" translation units, since they can not be considered as translation equivalents outside of this sentence context. The reason for is that e*ines Mordes* is the NP with indefinite article in Genitive case, and thus, according to the TIGER Annotation Scheme, it is the Genitive Attribute. Its Georgian equivalent is a noun with an enclitic postposition particle without any semantic feature of definiteness. Another NP *dem Verdacht* ("the suspicion") with definite article in Dative case, is linked to the Georgian parallel token in Genitive case ეჭვის followed by გამო ("due to", "because of") postposition that stands on its own.

Although, in the sketched case we have a conventional rather than a translation equivalence, the PP and PSP are spanned as a "good" alignment.

The reason of linking the NP *Die Polizei* ("die police") to the token პოლიციამ as "fuzzy" alignment is the morphological marker of the Ergative case მ ("-m") in the end, whereas the German NP is in the Nominative. Duo to the typological divergence, already mentioned, the Georgian nouns lack the "definiteness/indefiniteness" feature. This point could be ignored as it is a structural difference and, hence, a redundant condition for the alignment issue. But another argument preventing to count them as "good" equivalents is that NP *Die Polizei* ("die police") can be also aligned to the same Georgian token პოლიცია in the Nominative case. Moreover, it is also a translation of the German word *Polizei* ("police" as establishment) without an article.

The 4th token in the German clause, Object Accusative (OA), *ihn* ("him") is aligned to the 2nd token in the Georgian clause to a pronoun ის in Nominative that also might be spanned with the German pronoun *Er* outside of the current sentence context (cl. Figure 11).

Despite the outlined differences as the consequence of various reasons, the parallel tree from the Figure 13 is a rare example of a "good" alignment with almost all tokens spanned as "fuzzy" ones on the word order layer. Normally, these kind of bilingual sentences, should not yield to a parallel trees with the "good" alignment.

4 Conclusions

In the present paper we gave a short outline of our observations for inference of translation equivalents in parallel texts of languages with diverse structures - German and Georgian, an agglutinative language with a rich and productive morphology, relatively free word order and a small Treebank

We discussed the typological differences manifested in diverse morphological and syntactic structures and tried to visualize them in the presented examples.

Alongside with the main principal for alignment of tokens and phrases in parallel text that employs "translation equivalents outside of the context", we have discussed also an approach that advocates a

different method – "a conventional translation on the pragmatic level". The later could be extended beyond the scope of the German-Georgian treebank building issue and applied in general to the bilingual parallel text alignment procedures for typologically dissimilar languages.

References

1. Brants, S. and Hansen, S.: 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), Las Palmas, pp. 1643–1649 (2002)
2. Breindl, E.: Präpositionalphrasen. In: Agel, Vilmos, Eichinger, Ludwig M., Eroms, Hans-Werner, Hellwig, Peter, Heringer, Hans Jürgen, Lobin, Henning (Hrsg.): Dependenz und Valenz. Ein internationales Handbuch der Zeitgenössischen Forschung. 2, Halbband, Berlin, New York: De Gruyter, pp. 936-951, (2006)
3. Grimes, S, Li, X., Bies, A., Kulick, S., Ma, X., Strassel, S. :Creating Arabic-English Parallel Word-Aligned Treebank Corpora at LDC. In 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)
4. Kapanadze O., Mishchenko, A.: *A Multilingual GRUG Treebank for Underresourced Languages.* In: A. Gelbukh (Ed): CICLing-2013: Computational Linguistics and Intelligent Text Processing, Part I, Lecture Notes in Computer Science, Vol. 7816, Springer-Verlag Berlin Heidelberg, pp. 50–59, (2013)
5. Kapanadze, O.: A German-Georgian Parallel Treebank Project. In Proceedings of the LREC2012 META-RESAERCH Workshop on Advanced Treebanking. Istanbul, Turkey (2012)
6. Kapanadze, O.: 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). Berlin Heidelberg: Springer-Verlag (2010)
7. Kapanadze, O.: 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,
1. South Africa (2009).
8. Klenk, U.: Generative Syntax. Tübingen: Günter Narr Verlag (2003).
9. Megyesi, B., & Dahlqvist, B.: A Turkish-Swedish Parallel Corpus and Tools for its Creation. In Proceedings of Nordiska Datalingvistdagarna (NoDaL- iDa 2007) (2007)

10. Megyesi, B., Hein Sa˚gvall, A., Csato´ Johanson, E.: Building a Swedish-Turkish Parallel Corpus. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)
11. Rios, A., Göhring, A., & Volk, M.: Quechua-Spanish Parallel Treebank. In 7th Conference on Treebanks and Linguistic Theories, Groningen (2009)
12. Samuelsson, Y., & Volk, M.: Presentation and Representation of Parallel Treebanks. In Proceedings of the Treebank - Workshop at Nodalida, Joensuu, Finland (2005)
13. Samuelsson, Y. and Volk, M.: Phrase Alignment in Parallel Treebanks. In Proceedings of 5th Workshop on Treebanks and Linguistic Theories Prague, Czech Republic (2006)
14. Synphaty: Syntax Editor – Manual – Nijmegen: Max Planck Institute for Psycholinguistics (2006)
15. Welke, K.: Einführung in die Satzanalyse: Die Bestimmung der Satzglieder im Deutschen. Berlin: Walter De Gruyter (2007)

OLEG KAPANADZE
TBILISI STATE UNIVERSITY,
TBILISI, GEORGIA
E-MAIL: <OK@CAUCASUS.NET>