ParGramBank: The ParGram Parallel Treebank

Sebastian Sulger and Miriam Butt

University of Konstanz, Germany

{sebastian.sulger|miriam.butt}@uni-konstanz.de

Tracy Holloway King eBay Inc., USA tracyking@ebay.com

Paul Meurer Uni Research AS, Norway paul.meurer@uni.no **Tibor Laczkó** and **György Rákosi** University of Debrecen, Hungary {laczko.tibor|rakosi.gyorgy}@arts.unideb.hu

Cheikh Bamba Dione and Helge Dyvik and Victoria Rosén and Koenraad De Smedt University of Bergen, Norway

dione.bamba@lle.uib.no, {dyvik|victoria|desmedt}@uib.no

Agnieszka Patejuk Polish Academy of Sciences aep@ipipan.waw.pl Özlem Çetinoğlu University of Stuttgart, Germany ozlem@ims.uni-stuttgart.de

I Wayan Arka^{*} and Meladel Mistica⁺ *Australian National University and Udayana University, Indonesia +Australian National University

wayan.arka@anu.edu.au, meladel.mistica@gmail.com

Abstract

This paper discusses the construction of a parallel treebank currently involving ten languages from six language families. The treebank is based on deep LFG (Lexical-Functional Grammar) grammars that were developed within the framework of the ParGram (Parallel Grammar) effort. The grammars produce output that is maximally parallelized across languages and language families. This output forms the basis of a parallel treebank covering a diverse set of phenomena. The treebank is publicly available via the INESS treebanking environment, which also allows for the alignment of language pairs. We thus present a unique, multilayered parallel treebank that represents more and different types of languages than are available in other treebanks, that represents

deep linguistic knowledge and that allows for the alignment of sentences at several levels: dependency structures, constituency structures and POS information.

1 Introduction

This paper discusses the construction of a parallel treebank currently involving ten languages that represent several different language families, including non-Indo-European. The treebank is based on the output of individual deep LFG (Lexical-Functional Grammar) grammars that were developed independently at different sites but within the overall framework of ParGram (the Parallel Grammar project) (Butt et al., 1999a; Butt et al., 2002). The aim of ParGram is to produce deep, wide coverage grammars for a variety of languages. Deep grammars provide detailed syntactic analysis, encode grammatical functions as well as

other grammatical features such as tense or aspect, and are linguistically well-motivated. The Par-Gram grammars are couched within the linguistic framework of LFG (Bresnan, 2001; Dalrymple, 2001) and are constructed with a set of grammatical features that have been commonly agreed upon within the ParGram group. ParGram grammars are implemented using XLE, an efficient, industrialstrength grammar development platform that includes a parser, a generator and a transfer system (Crouch et al., 2012). XLE has been developed in close collaboration with the ParGram project. Over the years, ParGram has continuously grown and includes grammars for Arabic, Chinese, English, French, German, Georgian, Hungarian, Indonesian, Irish, Japanese, Malagasy, Murrinh-Patha, Norwegian, Polish, Spanish, Tigrinya, Turkish, Urdu, Welsh and Wolof.

ParGram grammars produce output that has been parallelized maximally across languages according to a set of commonly agreed upon universal proto-type analyses and feature values. This output forms the basis of the ParGramBank parallel treebank discussed here. ParGramBank is constructed using an innovative alignment methodology developed in the XPAR project (Dyvik et al., 2009) in which grammar parallelism is presupposed to propagate alignment across different projections (section 6). This methodology has been implemented with a drag-and-drop interface as part of the LFG Parsebanker in the INESS infrastructure (Rosén et al., 2012; Rosén et al., 2009). ParGramBank has been constructed in INESS and is accessible in this infrastructure, which also offers powerful search and visualization.

In recent years, parallel treebanking¹ has gained in importance within NLP. An obvious application for parallel treebanking is machine translation, where treebank size is a deciding factor for whether a particular treebank can support a particular kind of research project. When conducting in-depth linguistic studies of typological features, other factors such as the number of included languages, the number of covered phenomena, and the depth of linguistic analysis become more important. The treebanking effort reported on in this paper supports work of the latter focus, including efforts at multilingual dependency parsing (Naseem et al., 2012). We have created a parallel treebank whose prototype includes ten typologically diverse languages and reflects a diverse set of phenomena. We thus present a unique, multilayered parallel treebank that represents more languages than are currently available in other treebanks, and different types of languages as well. It contains deep linguistic knowledge and allows for the parallel and simultaneous alignment of sentences at several levels. LFG's f(unctional)-structure encodes dependency structures as well as information that is equivalent to Quasi-Logical Forms (van Genabith and Crouch, 1996). LFG's c(onstituent)-structure provides information about constituency, hierarchical relations and part-of-speech. Currently, ParGramBank includes structures for the following languages (with the ISO 639-3 code and language family): English (eng, Indo-European), Georgian (kat, Kartvelian), German (deu, Indo-European), Hungarian (hun, Uralic), Indonesian (ind, Austronesian), Norwegian (Bokmål) (nob, Indo-European), Polish (pol, Indo-European), Turkish (tur, Altaic), Urdu (urd, Indo-European) and Wolof (wol, Niger-Congo). It is freely available for download under the CC-BY 3.0 license via the INESS treebanking environment and comes in two formats: a Prolog format and an XML format.²

This paper is structured as follows. Section 2 discusses related work in parallel treebanking. Section 3 presents ParGram and its approach to parallel treebanking. Section 4 focuses on the treebank design and its construction. Section 5 contains examples from the treebank, focusing on typological aspects and challenges for parallelism. Section 6 elaborates on the mechanisms for parallel alignment of the treebank.

2 Related Work

There have been several efforts in parallel treebanking across theories and annotation schemes.

Kuhn and Jellinghaus (2006) take a minimal approach towards multilingual parallel treebanking. They bootstrap phrasal alignments over a sentence-aligned parallel corpus of English, French, German and Spanish and report concrete treebank annotation work on a sample of sentences from the Europarl corpus. Their annotation

¹Throughout this paper 'treebank' refers to both phrasestructure resources and their natural extensions to dependency and other deep annotation banks.

²http://iness.uib.no. The treebank is in the public domain (CC-BY 3.0). The use of the INESS platform itself is not subject to any licensing. To access the treebank, click on 'Treebank selection' and choose the ParGram collection.

scheme is the "leanest" possible scheme in that it consists solely of a bracketing for a sentence in a language (where only those units that play the role of a semantic argument or modifier in a larger unit are bracketed) and a correspondence relation of the constituents across languages.

Klyueva and Mareček (2010) present a small parallel treebank using data and tools from two existing treebanks. They take a syntactically annotated gold standard text for one language and run an automated annotation on the parallel text for the other language. Manually annotated Russian data are taken from the SynTagRus treebank (Nivre et al., 2008), while tools for parsing the corresponding text in Czech are taken from the TectoMT framework (Popel and Žabokrtský, 2010).

The SMULTRON project is concerned with constructing a parallel treebank of English, German and Swedish. The sentences have been POS-tagged and annotated with phrase structure trees. These trees have been aligned on the sentence, phrase and word level. Additionally, the German and Swedish monolingual treebanks contain lemma information. The treebank is distributed in TIGER-XML format (Volk et al., 2010).

Megyesi et al. (2010) discuss a parallel English-Swedish-Turkish treebank. The sentences in each language are annotated morphologically and syntactically with automatic tools, aligned on the sentence and the word level and partially hand-corrected.³

A further parallel treebanking effort is Par-TUT, a parallel treebank (Sanguinetti and Bosco, 2011; Bosco et al., 2012) which provides dependency structures for Italian, English and French and which can be converted to a CCG (Combinatory Categorial Grammar) format.

Closest to our work is the ParDeepBank, which is engaged in the creation of a highly parallel treebank of English, Portuguese and Bulgarian. ParDeepBank is couched within the linguistic framework of HPSG (Head-Driven Phrase Structure Grammar) and uses parallel automatic HPSG grammars, employing the same tools and implementation strategies across languages (Flickinger et al., 2012). The parallel treebank is aligned on the sentence, phrase and word level.

In sum, parallel treebanks have so far focused exclusively on Indo-European languages (with Turkish providing the one exception) and generally do not extend beyond three or four languages. In contrast, our ParGramBank treebank currently includes ten typologically different languages from six different language families (Altaic, Austronesian, Indo-European, Kartvelian, Niger-Congo, Uralic).

A further point of comparison with ParDeep-Bank is that it relies on dynamic treebanks, which means that structures are subject to change during the further development of the resource grammars. In ParDeepBank, additional machinery is needed to ensure correct alignment on the phrase and word level (Flickinger et al., 2012, p. 105). ParGramBank contains finalized analyses, structures and features that were designed collaboratively over more than a decade, thus guaranteeing a high degree of stable parallelism. However, with the methodology developed within XPAR, alignments can easily be recomputed from f-structure alignments in case of grammar or feature changes, so that we also have the flexible capability of allowing ParGramBank to include dynamic treebanks.

3 ParGram and its Feature Space

The ParGram grammars use the LFG formalism which produces c(onstituent)-structures (trees) and f(unctional)-structures as the syntactic analysis. LFG assumes a version of Chomsky's Universal Grammar hypothesis, namely that all languages are structured by similar underlying principles (Chomsky, 1988; Chomsky, 1995). Within LFG, f-structures encode a language universal level of syntactic analysis, allowing for crosslinguistic parallelism at this level of abstraction. In contrast, c-structures encode language particular differences in linear word order, surface morphological vs. syntactic structures, and constituency (Dalrymple, 2001). Thus, while the Chomskyan framework is derivational in nature, LFG departs from this view by embracing a strictly representational approach to syntax.

ParGram tests the LFG formalism for its universality and coverage limitations to see how far parallelism can be maintained across languages. Where possible, analyses produced by the grammars for similar constructions in each language are parallel, with the computational advantage that the grammars can be used in similar applications and that machine translation can be simplified.

³The paper mentions Hindi as the fourth language, but this is not yet available: http://stp.lingfil.uu. se/~bea/turkiska/home-en.html.

The ParGram project regulates the features and values used in its grammars. Since its inception in 1996, ParGram has included a "feature committee", which collaboratively determines norms for the use and definition of a common multilingual feature and analysis space. Adherence to feature committee decisions is supported technically by a routine that checks the grammars for compatibility with a feature declaration (King et al., 2005); the feature space for each grammar is included in ParGramBank. ParGram also conducts regular meetings to discuss constructions, analyses and features.

For example, Figure 1 shows the c-structure of the Urdu sentence in (1) and the c-structure of its English translation. Figure 2 shows the fstructures for the same sentences. The left/upper c- and f-structures show the parse from the English ParGram grammar, the right/lower ones from Urdu ParGram grammar.^{4,5} The c-structures encode linear word order and constituency and thus look very different; e.g., the English structure is rather hierarchical while the Urdu structure is flat (Urdu is a free word-order language with no evidence for a VP; Butt (1995)). The f-structures, in contrast, are parallel aside from grammar-specific characteristics such as the absence of grammatical gender marking in English and the absence of articles in Urdu.⁶

 (1) کسان ني اپنا ٹريکٹر بيچا ؟
 kisAn=nE apnA farmer.M.Sg=Erg self.M.Sg TrEkTar bEc-A tractor.M.Sg sell-Perf.M.Sg
 'Did the farmer sell his tractor?'

With parallel analyses and parallel features, maximal parallelism across typologically different languages is maintained. As a result, during the construction of the treebank, post-processing and conversion efforts are kept to a minimum.

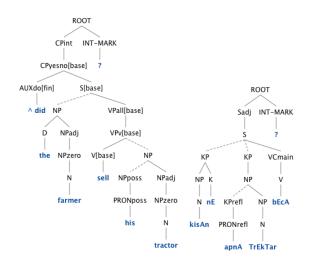


Figure 1: English and Urdu c-structures

We emphasize the fact that ParGramBank is characterized by a maximally reliable, humancontrolled and linguistically deep parallelism across aligned sentences. Generally, the result of automatic sentence alignment procedures are parallel corpora where the corresponding sentences normally have the same purported meaning as intended by the translator, but they do not necessarily match in terms of structural expression. In building ParGramBank, conscious attention is paid to maintaining semantic and constructional parallelism as much as possible. This design feature renders our treebank reliable in cases when the constructional parallelism is reduced even at fstructure. For example, typological variation in the presence or absence of finite passive constructions represents a case of potential mismatch. Hungarian, one of the treebank languages, has no productive finite passives. The most common strategy in translation is to use an active construction with a topicalized object, with no overt subject and with **3PL** verb agreement:

(2) A fá-t ki-vág-t-ák.the tree-ACC out-cut-PAST-3PL'The tree was cut down.'

In this case, a topicalized object in Hungarian has to be aligned with a (topical) subject in English. Given that both the sentence level and the phrase level alignments are human-controlled in the treebank (see sections 4 and 6), the greatest possible parallelism is reliably captured even in such cases of relative grammatical divergence.

⁴The Urdu ParGram grammar makes use of a transliteration scheme that abstracts away from the Arabic-based script; the transliteration scheme is detailed in Malik et al. (2010).

⁵In the c-structures, dotted lines indicate distinct functional domains; e.g., in Figure 1, the NP *the farmer* and the VP *sell his tractor* belong to different f-structures: the former maps onto the SUBJ f-structure, while the latter maps onto the topmost f-structure (Dyvik et al., 2009). Section 6 elaborates on functional domains.

⁶The CASE feature also varies: since English does not distinguish between accusative, dative, and other oblique cases, the OBJ is marked with a more general *obl* CASE.

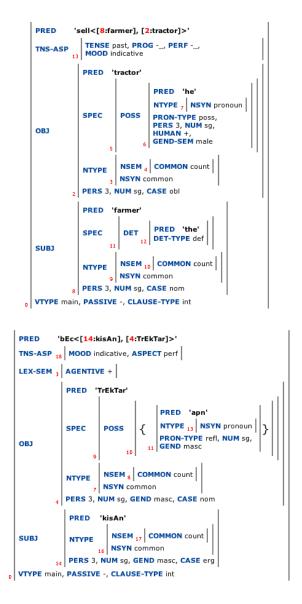


Figure 2: Parallel English and Urdu f-structures

4 Treebank Design and Construction

For the initial seeding of the treebank, we focused on 50 sentences which were constructed manually to cover a diverse range of phenomena (transitivity, voice alternations, interrogatives, embedded clauses, copula constructions, control/raising verbs, etc.). We followed Lehmann et al. (1996) and Bender et al. (2011) in using coverage of grammatical constructions as a key component for grammar development. (3) lists the first 16 sentences of the treebank. An expansion to 100 sentences is scheduled for next year.

- (3) a. Declaratives:
 - The driver starts the tractor.
 The tractor is red.

- b. Interrogatives:
 - 3. What did the farmer see?
 - 4. Did the farmer sell his tractor?
- c. Imperatives:
 - 5. Push the button.
 - 6. Don't push the button.
- d. Transitivity:7. The farmer gave his neighbor an old
 - tractor.
 - 8. The farmer cut the tree down.
 - 9. The farmer groaned.
- e. Passives and traditional voice:10. My neighbor was given an old tractor by the farmer.
 - 11. The tree was cut down yesterday.
 - 12. The tree had been cut down.
 - 13. The tractor starts with a shudder.
- f. Unaccusative:
 - 14. The tractor appeared.
- g. Subcategorized declaratives:
 - 15. The boy knows the tractor is red.
 - 16. The child thinks he started the tractor.

The sentences were translated from English into the other treebank languages. Currently, these languages are: English, Georgian, German, Hungarian, Indonesian, Norwegian (Bokmål), Polish, Turkish, Urdu and Wolof. The translations were done by ParGram grammar developers (i.e., expert linguists and native speakers).

The sentences were automatically parsed with ParGram grammars using XLE. Since the parsing was performed sentence by sentence, our resulting treebank is automatically aligned at the sentence level. The resulting c- and f-structures were banked in a database using the LFG Parsebanker (Rosén et al., 2009). The structures were disambiguated either prior to banking using XLE or during banking with the LFG Parsebanker and its discriminant-based disambiguation technique. The banked analyses can be exported and downloaded in a Prolog format using the LFG Parsebanker interface. Within XLE, we automatically convert the structures to a simple XML format and make these available via ParGramBank as well.

The Prolog format is used with applications which use XLE to manipulate the structures, e.g. for further semantic processing (Crouch and King, 2006) or for sentence condensation (Crouch et al., 2004).

5 Challenges for Parallelism

We detail some challenges in maintaining parallelism across typologically distinct languages.

5.1 Complex Predicates

Some languages in ParGramBank make extensive use of complex predicates. For example, Urdu uses a combination of predicates to express concepts that in languages like English are expressed with a single verb, e.g., 'memory do' = 'remember', 'fear come' = 'fear'. In addition, verb+verb combinations are used to express permissive or aspectual relations. The strategy within ParGram is to abstract away from the particular surface morphosyntactic expression and aim at parallelism at the level of f-structure. That is, monoclausal predications are analyzed via a simple f-structure whether they consist of periphrastically formed complex predicates (Urdu, Figure 3), a simple verb (English, Figure 4), or a morphologically derived form (Turkish, Figure 5).

In Urdu and in Turkish, the top-level PRED is complex, indicating a composed predicate. In Urdu, this reflects the noun-verb complex predicate *sTArT kar* 'start do', in Turkish it reflects a morphological causative. Despite this morphosyntactic complexity, the overall dependency structure corresponds to that of the English simple verb.

- (4) لارايئور ٹريکٹر کو سٹارٹ کرتا ہي
 (4) DrAIvar TrEkTar=kO
 (4) driver.M.Sg.Nom tractor.M.Sg=Acc
 (5) sTArT kartA hE
 (5) start.M.Sg do.Impf.M.Sg be.Pres.3Sg
 (5) 'The driver starts the tractor.'
- (5) sürücü traktör-ü çalış-tır-ıyor driver.Nom tractor-Acc work-Caus-Prog.3Sg 'The driver starts the tractor.'

The f-structure analysis of complex predicates is thus similar to that of languages which do not use complex predicates, resulting in a strong syntactic parallelism at this level, even across typologically diverse languages.

5.2 Negation

Negation also has varying morphosyntactic surface realizations. The languages in ParGramBank differ with respect to their negation strategies. Languages such as English and German use independent negation: they negate using words such as

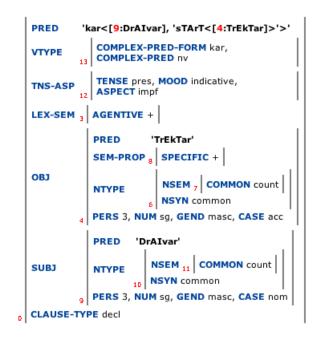


Figure 3: Complex predicate: Urdu analysis of (4)

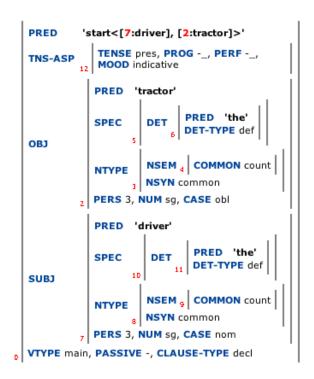


Figure 4: Simple predicate: English analysis of (4)

adverbs (English *not*, German *nicht*) or verbs (English *do*-support). Other languages employ nonindependent, morphological negation techniques; Turkish, for instance, uses an affix on the verb, as in (6).

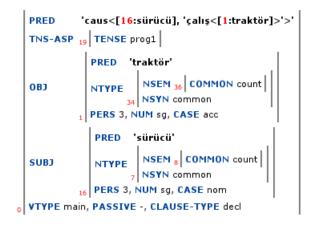


Figure 5: Causative: Turkish analysis of (5)

(6) düğme-ye bas-ma button-Dat push-Neg.Imp'Don't push the button.'

Within ParGram we have not abstracted away from this surface difference. The English *not* in (6) functions as an adverbial adjunct that modifies the main verb (see top part of Figure 6) and information would be lost if this were not represented at f-structure. However, the same cannot be said of the negative affix in Turkish — the morphological affix is not an adverbial adjunct. We have therefore currently analyzed morphological negation as adding a feature to the f-structure which marks the clause as negative, see bottom half of Figure 6.

5.3 Copula Constructions

Another challenge to parallelism comes from copula constructions. An approach advocating a uniform treatment of copulas crosslinguistically was advocated in the early years of ParGram (Butt et al., 1999b), but this analysis could not do justice to the typological variation found with copulas. Par-GramBank reflects the typological difference with three different analyses, with each language making a language-specific choice among the three possibilities that have been identified (Dalrymple et al., 2004; Nordlinger and Sadler, 2007; Attia, 2008; Sulger, 2011; Laczkó, 2012).

The possible analyses are demonstrated here with respect to the sentence *The tractor is red*. The English grammar (Figure 7) uses a raising approach that reflects the earliest treatments of copulas in LFG (Bresnan, 1982). The copula takes a non-finite complement whose subject is raised to the matrix clause as a non-thematic subject of the copula. In contrast, in Urdu (Figure 8), the

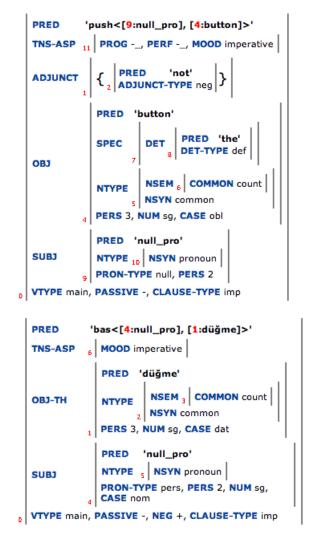


Figure 6: Different f-structural analyses for negation (English vs. Turkish)

copula is a two-place predicate, assigning SUBJ and PREDLINK functions. The PREDLINK function is interpreted as predicating something about the subject. Finally, in languages like Indonesian (Figure 9), there is no overt copula and the adjective is the main predicational element of the clause.

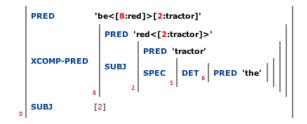


Figure 7: English copula example

	PRED '	'hO<[<mark>4</mark> :TrEkTar], [<mark>3</mark> :surx]>'	
	PREDLINK 3	PRED 'surx'	
D	SUBJ 4	PRED 'TrEkTar'	

Figure 8: Urdu copula example



Figure 9: Indonesian copula example

5.4 Summary

This section discussed some challenges for maintaining parallel analyses across typologically diverse languages. Another challenge we face is when no corresponding construction exists in a language, e.g. with impersonals as in the English *It is raining*. In this case, we provide a translation and an analysis of the structure of the corresponding translation, but note that the phenomenon being exemplified does not actually exist in the language. A further extension to the capabilities of the treebank could be the addition of pointers from the alternative structure used in the translation to the parallel aligned set of sentences that correspond to this alternative structure.

6 Linguistically Motivated Alignment

The treebank is automatically aligned on the sentence level, the top level of alignment within Par-GramBank. For phrase-level alignments, we use the drag-and-drop alignment tool in the LFG Parsebanker (Dyvik et al., 2009). The tool allows the alignment of f-structures by dragging the index of a subsidiary source f-structure onto the index of the corresponding target f-structure. Two f-structures correspond if they have translationally matching predicates, and the arguments of each predicate correspond to an argument or adjunct in the other f-structure. The tool automatically computes the alignment of c-structure nodes on the basis of the manually aligned corresponding f-structures.⁷

This method is possible because the c-structure to f-structure correspondence (the ϕ relation) is encoded in the ParGramBank structures, allowing the LFG Parsebanker tool to compute which cstructure nodes contributed to a given f-structure via the inverse (ϕ^{-1}) mapping. A set of nodes mapping to the same f-structure is called a 'functional domain'. Within a source and a target functional domain, two nodes are automatically aligned only if they dominate corresponding word forms. In Figure 10 the nodes in each functional domain in the trees are connected by whole lines while dotted lines connect different functional domains. Within a functional domain, thick whole lines connect the nodes that share alignment; for simplicity the alignment is only indicated for the top nodes. The automatically computed c-structural alignments are shown by the curved lines. The alignment information is stored as an additional layer and can be used to explore alignments at the string (word), phrase (c-)structure, and functional (f-)structure levels.

We have so far aligned the treebank pairs English-Urdu, English-German, English-Polish and Norwegian-Georgian. As Figure 10 illustrates for (7) in an English-Urdu pairing, the English object *neighbor* is aligned with the Urdu indirect object (OBJ-GO) *hamsAyA* 'neighbor', while the English indirect object (OBJ-TH) *tractor* is aligned with the Urdu object *TrEkTar* 'tractor'. The cstructure correspondences were computed automatically from the f-structure alignments.

(7) لني اپني ہمسايي کو پرانا ٹريکٹر ديا kisAn=nE apnE farmer.M.Sg=Erg self.Obl hamsAyE=kO purAnA neighbor.M.Sg.Obl=Acc old.M.Sg TrEkTar di-yA tractor.M.Sg give-Perf.M.Sg 'The farmer gave his neighbor an old tractor.'

The INESS platform additionally allows for the highlighting of connected nodes via a mouse-over technique. It thus provides a powerful and flexible tool for the semi-automatic alignment and subse-

⁷Currently we have not measured inter-annotator agreement (IAA) for the f-structure alignments. The f-structure alignments were done by only one person per language pair. We anticipate that multiple annotators will be needed for this

task in the future, in which case we will measure IAA for this step.

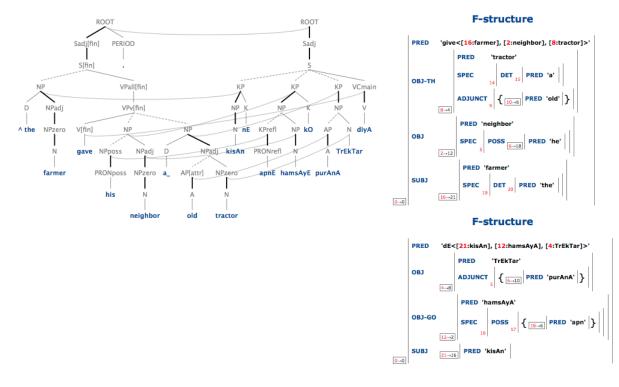


Figure 10: Phrase-aligned treebank example English-Urdu: The farmer gave his neighbor an old tractor.

quent inspection of parallel treebanks which contain highly complex linguistic structures.⁸

7 Discussion and Future Work

We have discussed the construction of ParGram-Bank, a parallel treebank for ten typologically different languages. The analyses in ParGram-Bank are the output of computational LFG Par-Gram grammars. As a result of ParGram's centrally agreed upon feature sets and prototypical analyses, the representations are not only deep in nature, but maximally parallel. The representations offer information about dependency relations as well as word order, constituency and part-ofspeech.

In future ParGramBank releases, we will provide more theory-neutral dependencies along with the LFG representations. This will take the form of triples (King et al., 2003). We also plan to provide a POS-tagged and a named entity marked up version of the sentences; these will be of use for more general NLP applications and for systems which use such markup as input to deeper processing. Third, the treebank will be expanded to include 100 more sentences within the next year. We also plan to include more languages as other ParGram groups contribute structures to ParGramBank.

ParGramBank, including its multilingual sentences and all annotations, is made freely available for research and commercial use under the CC-BY 3.0 license via the INESS platform, which supports alignment methodology developed in the XPAR project and provides search and visualization methods for parallel treebanks. We encourage the computational linguistics community to contribute further layers of annotation, including semantic (Crouch and King, 2006), abstract knowledge representational (Bobrow et al., 2007), Prop-Bank (Palmer et al., 2005), or TimeBank (Mani and Pustejovsky, 2004) annotations.

References

- Mohammed Attia. 2008. A Unified Analysis of Copula Constructions. In *Proceedings of the LFG '08 Conference*, pages 89–108. CSLI Publications.
- Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2011. Grammar Engineering and Linguistic Hypothesis Testing: Computational Support for Complexity in Syntactic Analysis. In Emily M. Bender and Jennifer E. Arnold, editors, *Languages from a Cognitive Perspective: Grammar, Usage and Processing*, pages 5–30. CSLI Publications.

⁸One reviewer inquires about possibilities of linking (semi-)automatically between languages, for example using lexical resources such as WordNets or Panlex. We agree that this would be desirable, but unrealizable, since many of the languages included in ParGramBank do not have a WordNet resource and are not likely to achieve an adequate one soon.

- Daniel G. Bobrow, Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Lottie Price, and Annie Zaenen. 2007. Precision-focused Textual Inference. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing.
- Cristina Bosco, Manuela Sanguinetti, and Leonardo Lesmo. 2012. The Parallel-TUT: a multilingual and multiformat treebank. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1932–1938, Istanbul, Turkey. European Language Resources Association (ELRA).
- Joan Bresnan. 1982. The Passive in Lexical Theory. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 3–86. The MIT Press.
- Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell Publishing.
- Miriam Butt, Stefanie Dipper, Anette Frank, and Tracy Holloway King. 1999a. Writing Large-Scale Parallel Grammars for English, French and German. In *Proceedings of the LFG99 Conference*. CSLI Publications.
- Miriam Butt, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. 1999b. A Grammar Writer's Cookbook. CSLI Publications.
- Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In *Proceedings of the COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.
- Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu*. CSLI Publications.
- Noam Chomsky. 1988. Lectures on Government and Binding: The Pisa Lectures. Foris Publications.
- Noam Chomsky. 1995. *The Minimalist Program*. MIT Press.
- Dick Crouch and Tracy Holloway King. 2006. Semantics via F-structure Rewriting. In *Proceedings of the LFG06 Conference*, pages 145–165. CSLI Publications.
- Dick Crouch, Tracy Holloway King, John T. Maxwell III, Stefan Riezler, and Annie Zaenen. 2004. Exploiting F-structure Input for Sentence Condensation. In *Proceedings of the LFG04 Conference*, pages 167–187. CSLI Publications.
- Dick Crouch, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula Newman, 2012. *XLE Documentation*. Palo Alto Research Center.
- Mary Dalrymple, Helge Dyvik, and Tracy Holloway King. 2004. Copular Complements: Closed or Open? In *Proceedings of the LFG '04 Conference*, pages 188–198. CSLI Publications.

- Mary Dalrymple. 2001. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press.
- Helge Dyvik, Paul Meurer, Victoria Rosén, and Koenraad De Smedt. 2009. Linguistically Motivated Parallel Parsebanks. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*, pages 71–82, Milan, Italy. EDU-Catt.
- Dan Flickinger, Valia Kordoni, Yi Zhang, António Branco, Kiril Simov, Petya Osenova, Catarina Carvalheiro, Francisco Costa, and Sérgio Castro. 2012. ParDeepBank: Multiple Parallel Deep Treebanking. In Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories (TLT11), pages 97–107, Lisbon. Edições Colibri.
- Tracy Holloway King, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ronald Kaplan. 2003. The PARC700 Dependency Bank. In *Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03).*
- Tracy Holloway King, Martin Forst, Jonas Kuhn, and Miriam Butt. 2005. The Feature Space in Parallel Grammar Writing. In Emily M. Bender, Dan Flickinger, Frederik Fouvry, and Melanie Siegel, editors, *Research on Language and Computation: Special Issue on Shared Representation in Multilingual Grammar Engineering*, volume 3, pages 139–163. Springer.
- Natalia Klyueva and David Mareček. 2010. Towards a Parallel Czech-Russian Dependency Treebank. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora*, Tartu. Northern European Association for Language Technology (NEALT).
- Jonas Kuhn and Michael Jellinghaus. 2006. Multilingual Parallel Treebanking: A Lean and Flexible Approach. In *Proceedings of the LREC 2006*, Genoa, Italy. ELRA/ELDA.
- Tibor Laczkó. 2012. On the (Un)Bearable Lightness of Being an LFG Style Copula in Hungarian. In *Proceedings of the LFG12 Conference*, pages 341–361. CSLI Publications.
- Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSNLP — Test Suites for Natural Language Processing. In Proceedings of COLING, pages 711–716.
- Muhammad Kamran Malik, Tafseer Ahmed, Sebastian Sulger, Tina Bögel, Atif Gulzar, Ghulam Raza, Sarmad Hussain, and Miriam Butt. 2010. Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta.

- Inderjeet Mani and James Pustejovsky. 2004. Temporal Discourse Models for Narrative Structure. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 57–64.
- Beáta Megyesi, Bengt Dahlqvist, Éva Á. Csató, and Joakim Nivre. 2010. The English-Swedish-Turkish Parallel Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective Sharing for Multilingual Dependency Parsing. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 629–637, Jeju Island, Korea, July. Association for Computational Linguistics.
- Joakim Nivre, Igor Boguslavsky, and Leonid Iomdin. 2008. Parsing the SynTagRus Treebank. In *Proceedings of COLING08*, pages 641–648.
- Rachel Nordlinger and Louisa Sadler. 2007. Verbless Clauses: Revealing the Structure within. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning, editors, *Architectures, Rules and Preferences: A Festschrift for Joan Bresnan*, pages 139–160. CSLI Publications.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.
- Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In *Proceedings* of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010), pages 293–304.
- Victoria Rosén, Paul Meurer, and Koenraad de Smedt. 2009. LFG Parsebanker: A Toolkit for Building and Searching a Treebank as a Parsed Corpus. In Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7), pages 127– 133, Utrecht. LOT.
- Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An Open Infrastructure for Advanced Treebanking. In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.
- Manuela Sanguinetti and Cristina Bosco. 2011. Building the Multilingual TUT Parallel Treebank. In *Proceedings of Recent Advances in Natural Language Processing*, pages 19–28.
- Sebastian Sulger. 2011. A Parallel Analysis of have-Type Copular Constructions in have-Less Indo-European Languages. In *Proceedings of the LFG* '11 Conference. CSLI Publications.

- Josef van Genabith and Dick Crouch. 1996. Direct and Underspecified Interpretations of LFG f-structures. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, volume 1, pages 262–267, Copenhagen, Denmark.
- Volk, Anne Göhring, Martin Torsten Marek, and Yvonne Samuelsson. 2010. SMUL-TRON (version 3.0) The Stockholm MULtilingual parallel TReebank. http://www.cl.uzh.ch/research/paralleltreebanks_en. html.