# Automatic Constituent and Function Alignment for Parallel Treebanking

26/09, 2010

**Abstract**

This paper describes the development of an automatic phrase alignment method using parallel sentences parsed in Lexical-Functional Grammar as input, where similarity in analyses is used as evidence that constituents or functional elements may be linked. A set of principles for phrase alignment are formulated, based on the goals of the XPar-project [1], and an implementation is given.

## 1 Introduction

Lexical-Functional Grammar (LFG) is a grammatical framework where a sentence is analysed as having both a constituent structure (c-structure) and functional structure (f-structure). The former is similar to traditional phrase structure trees, while the latter is an attribute-value matrix/graph which represents dependency relations between syntactic functions (subject, object, etc.), in addition to the grammatical features of these. The argument structure of predicates is embedded in the f-structure representation.

This work is part of the XPar-project, which involves developing a parallel treebank which will include links between corresponding constituents, as well as between corresponding syntactic functions. By utilising the information available in each monolingual LFG-parse of two parallel sentences, we are able to make precise and linguistically informative alignments on both the c-structure and f-structure level.

Although there exists many methods for automatic phrase alignment [3], most of these have been based on aligning any N-gram that is compatible with a word alignment, where none of these take into account syntactic features, and alignments may cross constituent borders. [2] describes a method for using statistical word-alignments as seeds to two separate constituent and dependency tree alignments; however, the goal here is to create a set of N-gram pairs for statistical machine translation, and the dependency and constituent alignments do not inform each other.

Our method is instead based on the fact that similar grammatical phenomena in different languages will have similar grammatical analyses, so structural similarity in the analyses should indicate that those parts of the analyses may be linked. How much structural similarity is required in order to link two elements is defined as a set of general constraints. This allows for a more top-down method of phrase alignment, which is more informative to the linguist since it links not only true constituents, but functional elements (which in LFG may even span discontiguous constituents). Word-alignments or translational dictionaries may be needed to automatically disambiguate in cases where the LFG parses do not give sufficient information; but the method will perform a large part of the alignment job even without *any* parallel corpus available.

The principles and constraints for alignment are discussed in the next section, section 3 describes the implementation, while section 4 discusses the strengths and weaknesses of the method.

## 2 Principles for Phrase Alignment

We want our alignment links to be useful for treebank studies, in the XPar-project this includes studying the relationship between syntactic function and semantic roles across languages, thus the principles for alignment (or, constraints on possible alignments) have to take this goal into account. An outline of the principles for phrase alignment used in the XPar-project are formulated in [1, pp. 75–77], this section recounts the major points, and explains some relevant LFG-terminology and concepts.

To introduce the relevant LFG-terminology, consider figure 1. This shows two simplified LFG f-structures and c-structures, ready for alignment. The English word *slept* is a verb phrase, and its nodes *project* the f-structure $g$ (as seen by the PRED value being the 'semantic form' of *slept*, '**sleep**'). The projection from c-structure to f-structure, $\phi$, is a many-to-one mapping, and all the nodes S, VP and V together project $g$. Since the nodes project the same f-structure, they constitute a *functional domain*. We can see that they project the same f-structure by the $\uparrow = \downarrow$ annotations, which are read as "my f-structure is the same as that of my mother node". The NP node has $\uparrow$ SUBJ $= \downarrow$ instead, read as "my f-structure is the SUBJ of my mother's f-structure"; the NP thus projects the value of the SUBJ f-structure inside $g$.

The argument structures of the Norwegian and English verbs are shown in their PRED values; both verbs take one argument, in the figure this is represented by an index. By looking up this index, we find that the one argument of '**sove**' is the subject of $f$, with '**eg**' as its PRED; similarly '**I**', subject of $g$, is the only argument of '**sleep**'. Neither of these subjects take any arguments themselves.

The candidates we consider for alignment are c-structure phrases, individual words, and PRED elements of f-structures[1]. In figure 1, we can link the PRED el-

[1]We could consider aligning other f-structure elements, but only PRED elements are sure to exist

ements of $f$ and $g$; by doing this we consider their f-structures linked. The PRED values of their arguments are also candidates for alignment, and in this case there would be no reason not to link them. As noted, the S, VP and V nodes in English constitute the functional domain of $g$, similarly IP, I' and V are the functional domain of $f$. Since their f-structures are linked, we have reason to link nodes from these functional domains. But we only want to link nodes if the material they dominate also corresponds: we would not want to link IP and S if the NP in Norwegian was linked to something that was not dominated by the S in English (or vice versa), since a c-structure link means that what is dominated by the linked nodes corresponds[2]. However, translations often ommit or add material, so an *unlinked* subordinate node (e.g. an adverbial only expressed in one language) should not interfere with the linking of IP and S.

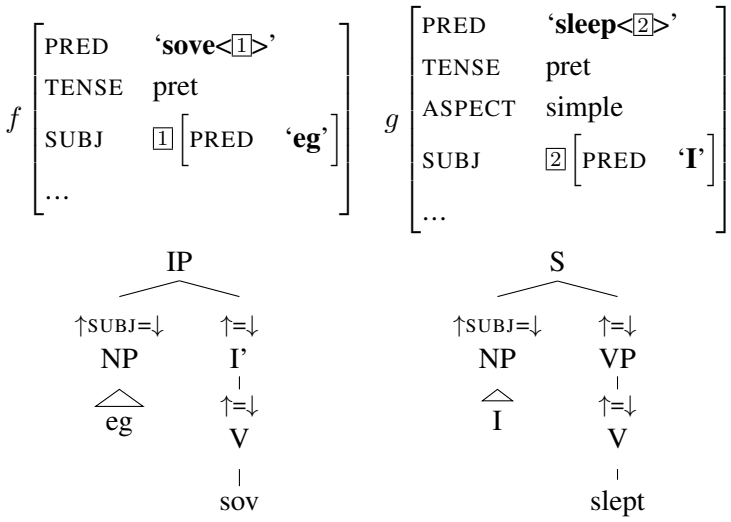


Figure 1: Example of simple links between constituents, f-structures and words (Norwegian and English)

Similarly, on the f-structure level we allow adjuncts (adverbials) to remain unlinked; adjuncts differ from arguments mainly in being non-obligatory, while arguments *are* required in order to express a certain sense of a predicate. So to link two predicates, we require all their arguments to find 'linguistically predictable translations' (LPT) in the translation, where a source word $W_s$ is LPT-correspondent with a target word $W_t$ if "$W_t$ can in general (out of context) be taken to be among the semantically plausible translations of $W_s$" [1, p. 74]. Nouns and pronominal forms are also considered LPT-correspondent.

in both languages, while grammatical features such as ASPECT$_{\{\}}$ might not exist in both languages, or be possible to link in a one-to-one-manner.

[2]Even if IP and S could not be linked, we could still link I' and VP, as these dominate the same linked material.

The argument structure of predicates in LFG is ordered, and this order typically reflects the semantic role hierarchy (agents being before themes, etc.). However, we do not require that linked arguments occupy the same positions in the argument structure of their predicates, since an English grammar may assign the first argument of the verb *like* to the agent, while a Spanish grammar may assign the first argument of the translation, *gustar*, to the theme. As one of the goals of the XPar-project is to study the relationship between semantic role and syntactic function, the aligner cannot presume that the relationship always is straightforward.

If any of the arguments of two otherwise linkable predicates do not have LPT-correspondents among each other, we have evidence that the predicates themselves are used to express different propositions. But should we allow adjuncts as translations of arguments? The examples in (1) are all translations of the same sentence; for the four different different languages, the grammar writers chose four different ways of dividing the participants in the verbal situation into arguments and adjuncts[3]. but in this translation, the predicates clearly express the same proposition. Thus we have to allow linking arguments to adjuncts; the monolingual evidence which informed the individual grammars may have suggested that a certain participant of a verbal situation should be analysed as an argument in one language, but as an adjunct in the other – in a particular translation, however, they may still correspond semantically.

(1) a. Adams veddet en sigarett med Browne på at det regnet. (Norwegian Bokmål)

$$\begin{bmatrix} \text{PRED} & \text{`}\textbf{vedde}\text{<Abrams, cigarette, Browne, rain>'} \\ \text{ADJUNCT} & \{\} \end{bmatrix}$$

b. abramsi brouns daenajleva sigaretze, rom cvimda. (Georgian)

$$\begin{bmatrix} \text{PRED} & \text{`}\textbf{da-najleveba}\text{<Abrams, Browne, regne>'} \\ \text{ADJUNCT} & \{\text{cigarette}\} \end{bmatrix}$$

c. Abrams hat mit Browne um eine Zigarette gewettet, daß es regnet. (German)

$$\begin{bmatrix} \text{PRED} & \text{`}\textbf{wetten}\text{<Abrams, regne>'} \\ \text{ADJUNCT} & \{\text{Browne, cigarette}\} \end{bmatrix}$$

d. Abrams bet a cigarette with Brown that it was raining. (English)

$$\begin{bmatrix} \text{PRED} & \text{`}\textbf{bet}\text{<Abrams, sigarett, regne>'} \\ \text{ADJUNCT} & \{\text{Browne}\} \end{bmatrix}$$

[3]The f-structures here are highly simplified.

More formally, these are the requirements for linking two f-structure PRED elements $p$ and $q$:

(2) a. the word-forms of $p$ and $q$ have LPT-correspondence
b. all arguments of $p$ have LPT-correspondence with an argument or adjunct of $q$
c. all arguments of $q$ have LPT-correspondence with an argument or adjunct of $p$
d. the LPT-correspondences are one-to-one
e. no adjuncts of $p$ are linked to f-structures outside $q$ or vice versa

The one-to-one requirement (2-d) is there to avoid linking two near-synonyms in one language into one word in the other language. We require all arguments of $p$ to have possible translations among the arguments and adjuncts of $q$, but we do not require (2) to be true of each argument of $p$; that is, an argument of $p$ may remain unlinked on the f-structure level. As mentioned, for adjuncts of $p$ we do not even require that they have LPT-correspondence with arguments/adjuncts of $q$, or vice versa, but (2-e) ensures that they are not *linked* outside of their predicates, which would imply that $p$ and $q$ did not contain corresponding linked material.

In order to link two c-structure nodes, [1, p. 77] defines the term *linked lexical nodes*, $LL$, where $LL(n)$ is the set of nodes dominated by $n$ which are word-linked. To link $n_s$ and $n_t$ (whose projected f-structures must be linked), all nodes in $LL(n_s)$ must be linked to nodes in $LL(n_t)$. Unlinked nodes dominated by $n_s$ or $n_t$ are not an obstacle to linking these nodes. Thus in in figure 1, if the NP nodes are linked, we may link IP and S, while in figure 2, the Norwegian I' and lower Georgian IP node may not be linked since the IP node dominates *robotebze*, linked to *roboter*, which is outside the nodes dominated by I'[4]. Georgian being a pro-drop language, the argument expressed by *de* in Norwegian does not have to be overtly expressed in Georgian, so there is no c-structure link for this word[5]. But by the criterion above we can still link the upper IP nodes, as they dominate the same sets of linked lexical nodes; the adjunct *gzaSi* ("on the way") is a translators addition only seen in the Georgian text, and remains unlinked both on c-structure and f-structure level, it does not stop linking the IP nodes.

By the above criterion, we may also link the Norwegian VP and Georgian I' nodes, since they dominate the same linked lexical nodes, *laparakobdnen* and *snakket*. However, *laparakobdnen* specifies a non-overt third person plural subject, while *snakket* does not. On the f-structure level, this pro-subject is linked to the Norwegian subject (*de* in the c-structure); a treebank user may want to exclude the link between the VP and I' nodes because of this discrepancy. Formally, we can exclude this kind of link by adding any linked f-structure arguments (of the

[4]The notation $\downarrow \in \uparrow$ ADJUNCT reads "my f-structure is a member of the set of adjuncts of my mother's f-structure" (a predicate may have only one subject, but an arbitrary number of adjuncts). Figure 2 is another example of phrases analysed as adjuncts in one language corresponding to phrases analysed as arguments in another language.

[5]The pro-subjects will be linked in f-structure, however.

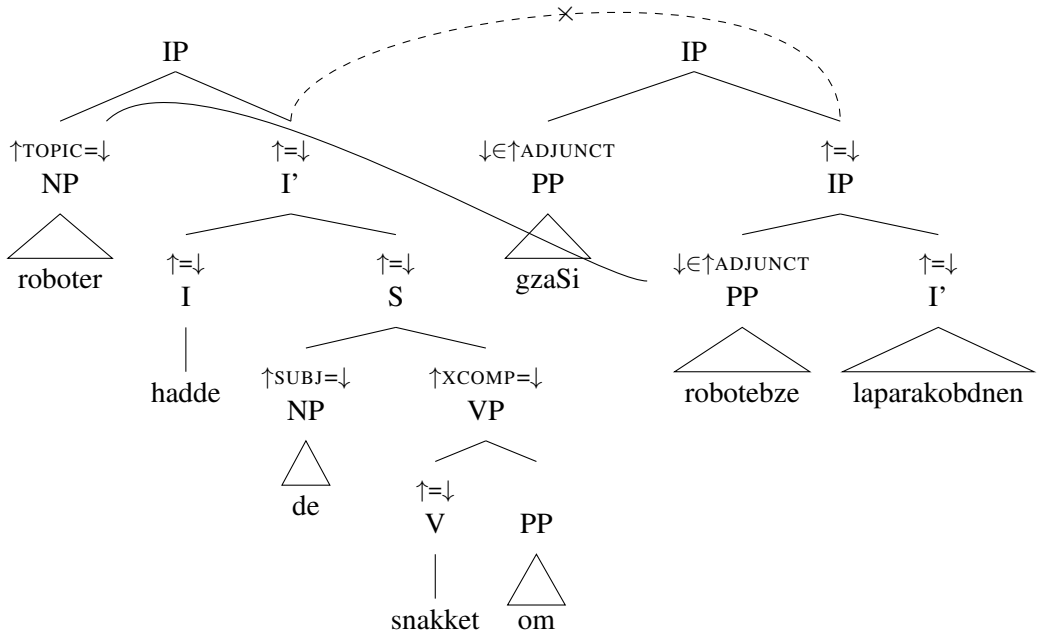


Figure 2: C-structure links must dominate the same set of links (Norwegian Bokmål "robots, had they talked about" and Georgian "on.the.way, about.robots they.had.talked")

f-structure projected by $n$) that are not overtly expressed, to \$LL(n)\$[6].

Section 4 notes some outstanding challenges with the linking principles, while the next section discusses the current implementation.

# 3 Implementation

This section covers a work-in-progress implementation of the above alignment principles[7].

# 4 Discussion

## 4.1 Challenges with the current implementation

Currently, we have no simple way of dealing with many-to-one alignments of PRED-elements. Figure 2 hides this problem; the current XPar grammars analyse *laparakobdnen* (they.had.talked) as a single predicate, while treating *hadde* (the perfective auxiliary) and *snakket* (talked) as two separate predicates. One might argue that then such phenomena should be analysed similarly, but as it is the goal of the aligner to help in discovering cross-language differences, grammars cannot be changed just to make the alignment easier.

# 5 Conclusion

# References


[1] Helge Dyvik, Paul Meurer, Victoria Rosén, and Koenraad De Smedt. Linguistically motivated parallel parsebanks. In Marco Passarotti, Adam Przepiórkowski, Savina Raynaud, and Frank Van Eynde, editors, *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*, pages 71–82, Milan, Italy, 2009. EDUCatt.

[2] M. Hearne, S. Ozdowska, and J. Tinsley. Comparing Constituency and Dependency Representations for SMT Phrase-Extraction. In *Actes de la 15e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN '08)*, Avignon, France, 2008.

[3] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.


[6] We cannot add just any *overtly* expressed argument to $LL$, as that would let us link the Norwegian I' and the Georgian IP node.

[7] All code available from http://example.com under the GNU General Public License, version 2 or later, along with some examples of input parses.