Syntax-based Constituent and Function Alignment for Parallel Treebanking

28/09, 2010

Abstract

This paper describes the development of an automatic phrase alignment method using as input parallel sentences parsed in Lexical-Functional Grammar, where similarity in analyses is used as evidence that constituents (syntactic phrases) or functional elements (predicates, arguments, adjuncts) may be linked. A set of principles for phrase alignment are formulated, with the goal of creating a parallel treebank for linguistic research, 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 functional relations between constituents (predicates and their subjects, objects, etc.), in addition to the grammatical features of these. The argument structure of predicates is embedded in the f-structure representation.

The work presented here is part of a master's thesis using resources from the XPar-project [2], which involves developing an LFG-parsed parallel treebank for Dutch, Tigrinya, Georgian and Norwegian, 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 in this treebank, we aim to create precise and linguistically informative alignments on both the c-structure and f-structure level.

Although there exist many methods for automatic phrase alignment, e.g. [4], most of these have been based on aligning any N-gram that is compatible with a word alignment, where syntactic features are not taken into account, and alignments may cross constituent borders. Later work has used statistical word-alignments as seeds to both constituent and dependency tree alignments, e.g. [3], but the separate dependency and constituent alignments created here do not inform each other.

Additionally, the goal has often been to create a set of N-gram pairs for statistical machine translation rather than a linguistically informative treebank; however, there has been newer research converting the output of these N-gram-based alignments into treebanks suitable for linguistic research [6].

Our method is instead based on the idea that similar grammatical phenomena in different languages will, if the grammars are correct, be given similar grammatical analyses[1], so structural similarity in the analyses indicates that those parts of the analyses may be linked. How much structural similarity we require in order to link two elements is defined as a set of general, language-independent constraints. This allows for a more top-down method of phrase alignment, the results of which are highly informative to the treebank user since we get links not only between true constituents, but between functional elements: predicates, arguments and adjuncts[2].

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 apart from the sentences to be aligned.

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 has already been formulated [2, pp. 75–77], this section recounts the major points while also delving into some corner cases, and explains the relevant LFG-terminology and concepts.

To introduce the relevant LFG-terminology, consider figure 3. 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, ϕ, 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

[1] Analysing similar phenomena in similar ways is a central guideline for grammar writers in the XPar-project, as well as of the overarching ParGram-project [1].

[2] In LFG these functional elements may even span discontiguous constituents.

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[3]. In figure 3, we can link the PRED elements 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[4]. 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.

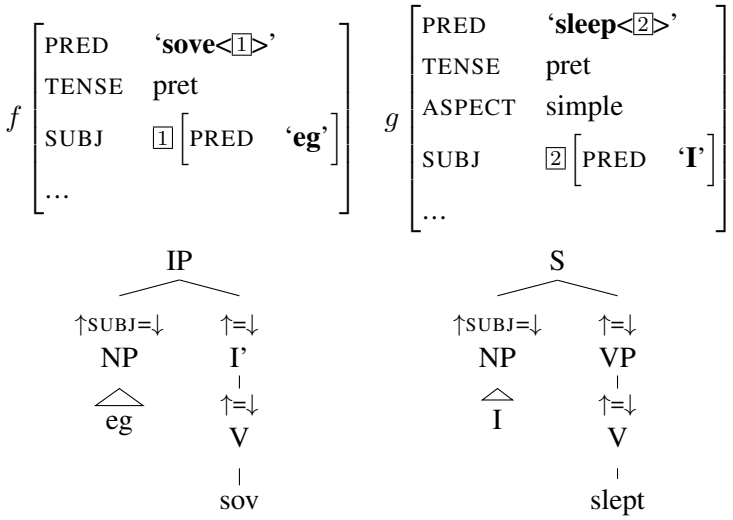

Figure 1: Example of simple links between constituents, f-structures and words (Norwegian and English)

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

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

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" [2, p. 74]. Nouns and pronominal forms are also considered LPT-correspondent.

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. However, given insufficient information, similarity in order may be used to *rank* different possible f-structure alignments.

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[5]. 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)

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

[5]The f-structures here are highly simplified, the analyses come from the grammars of the ParGram-project [1].

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

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

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

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

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

Additionally, when an argument/adjunct is selected by a preposition we skip the PRED of the preposition and consider its object as if there were no preposition there.

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, [2, 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 3, if the NP nodes are linked, we may link IP and S.

Figure 2 shows a much more complex situation, here 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'[6]. Georgian being a

[6]The notation $\downarrow\in\uparrow$ ADJUNCT reads "my f-structure is a member of the set of adjuncts of my

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[7]. 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.

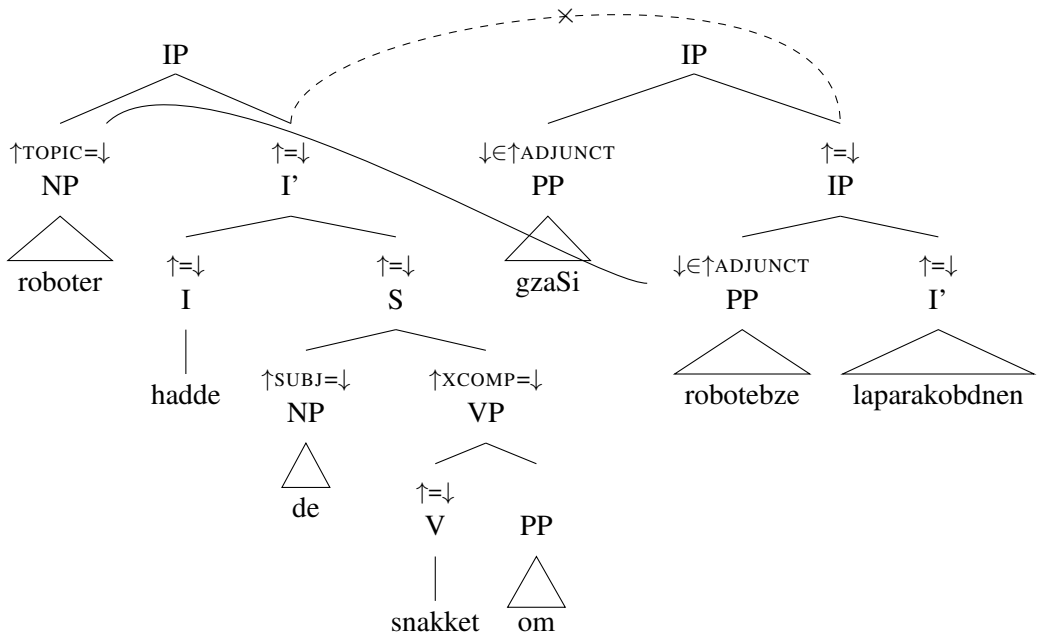

Figure 2: C-structure links must dominate the same set of links (Norwegian Bokmål "robots, had they talked about" and Georgian "on.the.way, about.robots they.had.talked")

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 f-structure projected by n) that are not overtly expressed, to \$LL(n)\$[8].

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.

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

[8]We cannot add just any *overtly* expressed argument to LL, as that would let us link the Norwe-

Several nodes may have equal LL, thus the c-structure links are often *many-to-many*. In addition, the f-structure PRED links are not always one-to-one, but this is a slightly more complex situation.

The f-structures of figure 2 need a many-to-many PRED link from *hadde* and *snakket* to *laparakobdnen*, since 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, all the while assuming that similar grammatical phenomena have similar grammatical analyses, grammars cannot be changed just to make the alignment easier – we have to treat this as a many-to-one PRED link[9].

In order to many-to-one-link p with q and a_q on the f-structure level, where a_q is an argument of q, the same requirements as (2) need to be fulfilled, but with the following difference: the argument lists of q and a_q are merged (as are their adjunct lists), with a_q not appearing in this list.

$$q\begin{bmatrix}\text{PRED} & \text{`}\mathbf{perf}\langle\boxed{1}\rangle\boxed{2}\text{'} \\ \text{SUBJ} & \boxed{2} \\ \text{TOPIC} & \boxed{3} \\ \text{XCOMP} & \boxed{1}\begin{bmatrix}\text{PRED} & \text{`}\mathbf{snakke*om}\langle\boxed{2},\boxed{3}\rangle\text{'} \\ \text{SUBJ} & \boxed{2}\begin{bmatrix}\text{PRED} & \text{`}\mathbf{de}\text{'}\end{bmatrix} \\ \text{OBJ} & \boxed{3}\begin{bmatrix}\text{PRED} & \text{`}\mathbf{robot}\text{'}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

$$p\begin{bmatrix}\text{PRED} & \text{`}\mathbf{laparaki}\langle\boxed{4}\rangle\text{'} \\ \text{SUBJ} & \boxed{4}\begin{bmatrix}\text{PRED} & \text{`}\mathbf{pro}\text{'}\end{bmatrix} \\ \text{ADJUNCT} & \left\{\begin{bmatrix}\text{PRED} & \text{`}\mathbf{Si}\langle\boxed{5}\rangle\text{'} \\ \text{OBJ} & \boxed{5}\begin{bmatrix}\text{PRED} & \text{`}\mathbf{gza}\text{'}\end{bmatrix}\end{bmatrix},\begin{bmatrix}\text{PRED} & \text{`}\mathbf{ze}\langle\boxed{6}\rangle\text{'} \\ \text{OBJ} & \boxed{6}\begin{bmatrix}\text{PRED} & \text{`}\mathbf{roboti}\text{'}\end{bmatrix}\end{bmatrix}\right\}\end{bmatrix}$$

Figure 3: Example of many-to-one link in f-structure: **perf** and **snakke*om** together link to **laparaki**.

So when attempting to link *hadde* (q) and *snakket* (a_q) with *laparakobdnen* (p), we merge the argument lists of q and its XCOMP argument, excluding the

gian I' and the Georgian IP node.

[9]Although in this case we might be able to align only the content verbs *hadde* and *laparakobdnen* by simply excluding auxiliary verbs from f-structure alignment, as with prepositions, there are other situations where we cannot avoid many-to-many links in a non-arbitrary fashion, e.g. lexical causatives linking to periphrastic causatives, argument incorporation, etc.

XCOMP itself, i.e. $\{\boxed{1},\boxed{2}\} \bigcup \{\boxed{2},\boxed{3}\} - \{\boxed{1}\} = \{\boxed{2},\boxed{3}\}$ (there are no adjuncts on the Norwegian side). Now we can link *laparakobdnen* with *hadde* and *snakket* by matching *de* ($\boxed{2}$) with the pro-element ($\boxed{4}$), and *robot* ($\boxed{3}$) with *roboti* ($\boxed{6}$).

The next section discusses the current implementation of these principles, while section 4 compares the possible merits of this method with other alignment methods.

3 Implementation

This section covers a work-in-progress implementation of the above alignment principles[10]. The program takes as input LFG-analyses of two sentences which we have for independent reasons consider as translations of each other. The analyses must be disambiguated and in the Prolog-format from XLE[11]. One may in addition give the program information about which word-translations are considered LPT, perhaps from automatic word-alignments or simple translational dictionaries.

The program begins by linking f-structures, where an f-structure *alignment* is a set of *links* between individual f-structures. The result of linking on this level may be ambiguous; since there are often several ways of linking arguments and adjuncts given insufficient LPT-information, we may end up with several possible f-structure alignments.

Therefore we rank the f-structure alignments. There are several possible ranking criteria, as mentioned above we use similarity in order of arguments to rank different possible f-structure alignments, when the LPT-information is not sufficient.

A single f-structure alignment is sent to the c-structure aligner, which by following the principles above always finds a single, unambiguous c-structure alignment (the different possible ways of calculating LL noted above are considered a user-option).

The f-structure aligner starts with the two outermost f-structures projected by LPT-correspondent words, and finds all possible ways of matching all arguments of the source PRED with LPT-correspondent arguments/adjuncts of the target PRED and vice versa (additionally adding any pairs of LPT-correspondent adjuncts that were not matched to arguments). For each of these possibilities, we recursively try to align the matched arguments/adjuncts[12], storing these possible sub-alignments in a table since solutions may overlap.

If we find no possibility of f-structure alignment (no way of fulfilling the requirements in (2) for the given PRED elements), we may try many-to-one links by merging argument lists as discussed in the previous section. Since this is not tried

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

[11] http://www2.parc.com/isl/groups/nltt/xle/doc/xle.html

[12] We allow PRED elements p and q to be linked even though some of their arguments cannot be recursively PRED-linked, as long as the requirement for word-level LPT-correspondence is fulfilled.

until there are no other possibilities, solutions involving many-to-one links of PRED elements are implicitly ranked lower than those where we can assume that translations corresponded better (a natural assumption since the sentences were aligned in the first place).

After ranking, finding the c-structure alignment for a single f-structure alignment is a simple matter of finding the LL for each node (being the union of the LL of each daughter node), and creating many-to-many links between those nodes that have the same LL. The many-to-many links here are the constituent alignment.

4 Discussion and outlook

The current implementation is, as mentioned, a work in progress (in particular, it does not yet have full support for f-structure links that are not one-to-one), which makes it difficult to do a complete evaluation with any statistical weight at this point. However, conducting tests on a set of example sentences chosen to illustrate a wide variety of grammatical phenomena, the results do seem promising.

Of course, the alignments will only be as good as the grammatical analyses that gave rise to them, so this is an important possible source of errors. Building high-quality, wide-coverage LFG grammars requires manual work that could be avoided if a large enough corpus is available.

However, a top-down method of alignment may be quite useful for lesser-resourced language pairs, where there exist LFG grammars for the languages. For a language pair such as Norwegian-Georgian or Tigrinya-Dutch, it is difficult to obtain a parallel corpus large enough to create high quality phrase alignments by purely corpus-based methods, not only because of the marginalisation of the languages, but also because of the productive morphology of Georgian. But by taking advantage of structural similarity in the LFG analyses of parallel sentences, the need for huge corpora is lessened[13]. Given some manual intervention in selecting between ambiguous alignments, not even a translational dictionary is needed, although this requires a suitable interface for manual selection of possible alignments[14].

References

[1] M. Butt, H. Dyvik, T.H. King, H. Masuichi, and C. Rohrer. The Parallel Grammar Project. In *COLING-02 on Grammar engineering and evaluation*, volume 15, pages 1–7, Morristown, NJ, 2002. International Conference on Computational Linguistics, Association for Computational Linguistics.

[13] Another issue is that using N-gram alignments created from corpora outside the domain of the treebank text (e.g. in order to increase recall) may hurt precision severely [6].

[14] The interface developed in [5] is in the process of being extended for alignment selection.

[2] 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.

[3] 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.

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

[5] Victoria Rosén, Paul Meurer, and Koenraad de Smedt. LFG Parsebanker: A Toolkit for Building and Searching a Treebank as a Parsed Corpus. In Frank Van Eynde, Anette Frank, Gertjan van Noord, and Koenraad De Smedt, editors, *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht, 2009. LOT.

[6] Y. Samuelsson and M. Volk. Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment. In *Proceedings of Treebanks and Linguistic Theories (TLT '07)*, Bergen, Norway, 2007.