LFG-based Constituent and Function Alignment for Parallel Treebanking

05/11, 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 annotating 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 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, the project aims 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 [5], 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 [4], 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 [9, 8, 3]. However, there has been newer research converting the output of these N-gram-based alignments into treebanks suitable for linguistic research [7].

The Xpar method is instead based on the idea that similar grammatical phenomena in different languages will, if the grammars are correct and constructed according to common principles, 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. In LFG these functional elements may even span discontinuous 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 apart from the sentences to be aligned.

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

2 Principles for Phrase Alignment

We want our alignments 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 that constraint possible alignments have to take this goal into account. An outline of the Xpar alignment principles has already been formulated [2, pp. 75–77]; this paper recounts the major points while also delving into some corner cases.

We begin by explaining the relevant LFG-terminology and concepts. Consider the Norwegian Nynorsk and English phrases in example (1) with analyses in figure 1. This shows two simplified LFG f-structures, with their c-structure trees below, ready for alignment. The English word *slept* is a verb phrase, and its nodes *project* the f-structure g (whose PRED value is the 'semantic form' of *slept*, '**sleep**'). The projection from c-structure to f-structure, ϕ, is a many-to-one mapping; all the nodes S, VP and V together project g. Since the nodes project the same f-structure, they constitute a *functional domain*. We see that they project the same f-structure by the $\uparrow{=}\downarrow$ annotations, which read "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 projects the value of SUBJ inside g.

[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], though in the latter only emphasising f-structure parallelism.

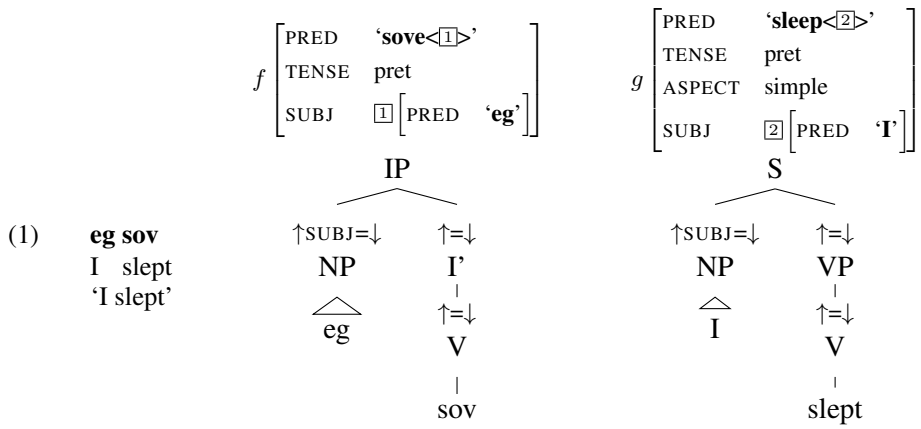

Figure 1: Example of simple linkable constituents, f-structures and words

The argument structures of the verbs are shown in their PRED values. Both take one argument; here represented by an index. Looking up the index, we find the one argument of '**sove**' is f's subject, with '**eg**' as its PRED. Similarly '**I**', g's subject, is the only argument of '**sleep**'. Neither subject takes any arguments itself.

Our alignment candidates are c-structure phrases, individual words, and PRED elements of f-structures.[2] In figure 1, we can link the PRED elements of f and g; by doing this we consider their f-structures linked. The PRED's of their arguments are also alignment candidates, 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[3]. However, translations often omit 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.

By the same logic, 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, the treebank guidelines 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,

[2]We could align other features, but only PRED's are sure to exist in both languages; grammatical features such as ASPECT might not exist in both languages, or be possible to link one-to-one.

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

p. 74]. Nouns and pronominal forms are also considered LPT-correspondent.

The argument structure of LFG predicates is ordered; the order typically reflects the semantic role hierarchy (agents before themes, etc.). However, we do not require that linked arguments occupy the same positions in the argument structure of their predicates. An English grammar may assign argument one of the verb *like* to the agent, while a Spanish grammar may assign argument one of *gustar* (a possible translation of *like*) to the theme. As a goal of the Xpar-project is to study the relationship between semantic role and syntactic function, the aligner cannot presume that the relationship is always straightforward. However, given insufficient information, similarity in order may be used to *rank* different possible alignment.

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 (2) are all translations of the same sentence, in English, Norwegian Bokmål, Georgian and German. 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.[4] But in this particular translation, the predicates clearly express the same proposition.

(2) a. **Abrams bet a cigarette with Browne that it was raining.**

$$\left[\text{PRED} \quad \text{`}\textbf{bet}\text{<Abrams, cigarette, rain>'} \quad \text{ADJUNCT} \quad \left\{\text{Browne}\right\}\right]$$

b. **Abrams veddet en sigarett med Browne på at det regnet.**
Abrams bet a cigarette with Browne on that it rained.

$$\left[\text{PRED} \quad \text{`}\textbf{bet}\text{<Abrams, cigarette, Browne, rain>'} \quad \text{ADJUNCT} \quad \left\{\ \right\}\right]$$

c. **abramsi brouns daenajleva sigaret-ze, rom cvimda.**
Abrams.NOM Browne.DAT bet.PERF cigarette.DAT-on, that rained.IMPERF.

$$\left[\text{PRED} \quad \text{`}\textbf{bet}\text{<Abrams, Browne, rain>'} \quad \text{ADJUNCT} \quad \left\{\text{cigarette}\right\}\right]$$

d. **Abrams hat mit Browne um eine Zigarette gewettet, daß es regnet.**
Abrams has with Browne about a cigarette.ACC bet, that it rained.

$$\left[\text{PRED} \quad \text{`}\textbf{bet}\text{<Abrams, rain>'} \quad \text{ADJUNCT} \quad \left\{\text{Browne, cigarette}\right\}\right]$$

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.

Note: in the f-structures above, some of the arguments/adjuncts are selected by prepositions, and their PRED will be embedded in the preposition's f-structure. In this situation, we skip the PRED of the preposition and consider its object as if there

[4] The PRED names in these f-structures have been translated to simplify the example. The analyses come from the grammars of the ParGram-project [1].

were no preposition there; this is necessary to align the participants in example (2).

Formally, to link two f-structure PRED elements p and q we require that all the following hold (see also [2]):

(3) 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 (skipping selectional prepositions)
c. all arguments of q have LPT-correspondence with an argument or adjunct of p (skipping selectional prepositions)
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 (3-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 (3) to be true recursively of each argument of p; that is, an argument of p may remain unlinked on the f-structure level. And for adjuncts of p we do not even require that they have LPT-correspondence with arguments/adjuncts of q, or vice versa, but (3-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 *word-linked* nodes[5] dominated by n. So:

(4) 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 figure 1, if the NP nodes are not linked to nodes outside these trees, we may link IP and S.

The Norwegian Bokmål and Georgian sentences in (5), with c-structures in figure 2, illustrate a much more complex situation.[6] Here the Norwegian I' and lower Georgian IP node may not be linked since the Georgian node dominates *roboṭebze*, linked to *roboter*, which is outside the nodes dominated by the I' node.[7]

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.[8] But by criterion (4) we can still link the upper IP nodes, as they dominate the same sets of linked lexical nodes. The adjunct *gzaši* is a translator's

[5] In the current implementation, word-links are defined by the PRED links of the f-structures they project.

[6] The sentences are from a book translation, but the Norwegian sentence has been topicalised to illustrate the c-structure constraint.

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

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

addition only seen in the Georgian text, and remains unlinked both on c-structure and f-structure level; it does not stop us from linking the IP nodes.

(5) a. **roboter hadde de snakket om**
robots had they talked about
'They had talked about *robots*'
b. **gza-ši roboṭeb-ze laparaḳobdnen**
way.DAT-to robots.DAT-on talked.3PL
'On the way, they had talked about robots'

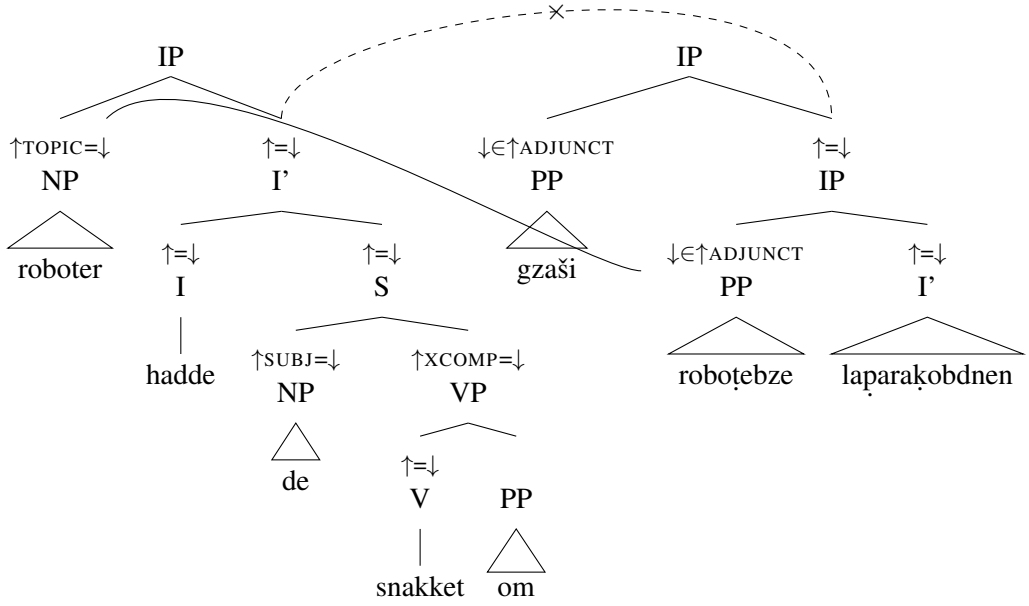

Figure 2: C-structure links must dominate the same set of links

By criterion (4), we may also link the Norwegian VP and Georgian I' nodes, since they dominate the same linked lexical nodes, *laparaḳobdnen* and *snakket*. However, *laparaḳobdnen* 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 to $LL(n)$ any linked f-structure arguments (of the f-structure projected by n) that are not overtly expressed.[9]

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

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

more involved problem. The f-structures of figure 3 need a many-to-one PRED link from '**perf**' and '**snakke*om**' to '**laṗaraḳi**', since the grammars analyse '**laṗaraḳi**' as a single predicate, while treating '**perf**' and '**snakke*om**' 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.[10]

$$p\begin{bmatrix}\text{PRED} & \text{`}\mathbf{perf}\langle\boxed{1}\rangle\boxed{2}\text{'}\\ \text{SUBJ} & \boxed{2}\\ \text{TOPIC} & \boxed{3}\\ \text{XCOMP} & \boxed{1}\begin{bmatrix}\text{PRED} & \text{`}\mathbf{snakke*om}\langle\boxed{2},\boxed{3}\rangle\text{'}\\ \text{SUBJ} & \boxed{2}\left[\text{PRED} \quad \text{`}\mathbf{de}\text{'}\right]\\ \text{OBJ} & \boxed{3}\left[\text{PRED} \quad \text{`}\mathbf{robot}\text{'}\right]\end{bmatrix}\end{bmatrix} \quad q\begin{bmatrix}\text{PRED} & \text{`}\mathbf{laṗaraḳi}\langle\boxed{4}\rangle\text{'}\\ \text{SUBJ} & \boxed{4}\left[\text{PRED} \quad \text{`}\mathbf{pro}\text{'}\right]\\ \text{ADJUNCT} & \left\{\begin{bmatrix}\text{PRED} & \text{`}\mathbf{ši}\langle\boxed{5}\rangle\text{'}\\ \text{OBJ} & \boxed{5}\left[\text{PRED} \quad \text{`}\mathbf{gza}\text{'}\right]\end{bmatrix},\right.\\ & \left.\begin{bmatrix}\text{PRED} & \text{`}\mathbf{ze}\langle\boxed{6}\rangle\text{'}\\ \text{OBJ} & \boxed{6}\left[\text{PRED} \quad \text{`}\mathbf{roboṭi}\text{'}\right]\end{bmatrix}\right\}\end{bmatrix}$$

Figure 3: F-structure many-to-one link from **perf** and **snakke*om** to **laṗaraḳi**.

In order to many-to-one-link from both p and a_p to q on the f-structure level, where a_p is an argument of p, the same requirements as in (3) need to be fulfilled, but with the following difference: the argument lists of p and a_p are merged (as are their adjunct lists), with a_p not appearing in this list.

So when attempting to link '**perf**' (p) and '**snakke*om**' (a_p) with '**laṗaraḳi**' (q), we merge the argument lists of p and its XCOMP argument, excluding the 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 '**laṗaraḳi**' with '**perf**' and '**snakke*om**' by matching '**de**' ($\boxed{2}$) with the pro-element ($\boxed{4}$), and '**robot**' ($\boxed{3}$) with '**roboṭi**' ($\boxed{6}$).

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

3 Implementation

This section covers a work-in-progress implementation of the above alignment principles.[11] The program takes as input LFG-analyses of two sentences which we consider as translations of each other for independent reasons. The analyses

[10] In this particular case we might be able to align only the content verbs *snakket* and *laṗaraḳobdnen* by excluding auxiliary verbs from f-structure alignment, as we do with prepositions. However, there are other situations where we cannot avoid non-one-to-one links in a non-arbitrary fashion, e.g. lexical causatives linking to periphrastic causatives, argument incorporation, etc.

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

must be disambiguated and in the XLE-format.[12] One may optionally supply information about which word-translations are considered LPT (e.g. from automatic word-alignments or 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.

The f-structure aligner, algorithm 1, starts with the two outermost f-structures projected by LPT-correspondent words. The helper $argalign$ returns all possible ways of matching all arguments of the source PRED with LPT-correspondent arguments/adjuncts of the target PRED and vice versa. For each of these possibilities, we recursively try to align the matched arguments/adjuncts,[13] storing these possible sub-alignments in a table since solutions may overlap.

$alignments \leftarrow \emptyset$;
forall the *argperm in argalign(F_s, F_t)* **do**
  $p \leftarrow \emptyset$;
  forall the A_s, A_t *in argperm* **do**
    if *unset(atab[A_s, A_t])* **then** atab[A_s, A_t] $\leftarrow$ f-align(A_s, A_t);
    $subalignment \leftarrow$ atab[A_s, A_t] ;
    if $subalignment$ **then** add $subalignment$ to p;
    else add (A_s, A_t) to p ; `// only LPT-correspondence`
  end
  add p to $alignments$;
  forall the *adjperm in adjalign(argperm, F_s, F_t)* **do**
    $d \leftarrow$ copy-of(p) ; `// optional adjunct links`
    forall the A_s, A_t *in adjperm* **do** as above, adding to d;
    add d to $alignments$;
  end
end
`// loop through adjalign if no arguments exist`
if $alignments = \emptyset$ **then** call f-align for each possible pred-arg merge ;
else return $((F_s, F_t), alignments)$;

Algorithm 1: f-align(F_s, F_t)

If we find no way of fulfilling the requirements in (3) for F_s and F_t, we may try many-to-one links by merging argument lists as discussed in the previous section. Since this is not tried 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).

Since f-align may give several solutions, we rank the f-alignments. There

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

[13] 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. Adjuncts not linked to arguments are optionally linked to each other.

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 of section 2 always finds a single, unambiguous c-structure alignment (the different possible ways of calculating LL noted above are considered a user-option). Finding the c-structure alignment for a single f-structure alignment involves first finding the LL for each node, where $LL(n)$ is the union of $LL(m)$ for all m dominated by of n; and then 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, making it difficult to do a complete evaluation at this point.[14] In particular: discontinuous constituents have not been fully tested, and the implementation currently uses bottom-up information as cut-offs instead of ranking. However, tests conducted on a set of example sentences, chosen to illustrate a wide variety of grammatical phenomena, 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 grammars requires manual work; however, without these, a large, informative and consistent treebank may require even more manual work.

A top-down method of alignment such as this may be quite useful for language pairs with few parallel resources, where there exist LFG grammars for the languages. For a language pair such as Norwegian-Georgian, it is difficult to obtain a parallel corpus large enough to create high quality phrase alignments purely by corpus-based methods, not only because of the marginality of the languages, but also because of the productive morphology of Georgian. By taking advantage of structural similarity in the LFG analyses of parallel sentences, the need for huge corpora is lessened.[15] Given some manual intervention in selecting between ambiguous alignments (and a suitable interface[16]), not even a translational dictionary is needed.

[14] Additionally, the program expects disambiguated analyses, and many sentences in the larger test sets have not seen manual disambiguation yet.

[15] Even where these are available, 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 [7, p. 149].

[16] The interface developed in [6] is in the process of being extended for alignment selection.

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, ACL.

[2] H. Dyvik, P. Meurer, V. Rosén, and K. De Smedt. Linguistically motivated parallel parsebanks. In M. Passarotti, A. Przepiórkowski, S. Raynaud, and F. Van Eynde, editors, *Proceedings of TLT8*, pages 71–82, Milan, Italy, 2009. EDUCatt.

[3] Yvette Graham, Anton Bryl, and Josef van Genabith. F-structure transfer-based statistical machine translation. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of LFG09*, pages 317–337, Trinity College, Cambridge, 2009. CSLI Publications.

[4] M. Hearne, S. Ozdowska, and J. Tinsley. Comparing Constituency and Dependency Representations for SMT Phrase-Extraction. In *TALN '08*, Avignon, France, 2008.

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

[6] V. Rosén, P. Meurer, and K. de Smedt. LFG Parsebanker: A Toolkit for Building and Searching a Treebank as a Parsed Corpus. In F. Van Eynde, A. Frank, G. van Noord, and K. De Smedt, editors, *Proceedings of TLT7*, pages 127–133, Utrecht, 2009. LOT.

[7] 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.

[8] J. Tiedemann and G. Kotzé. A discriminative approach to tree alignment. In *Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning*, pages 33–39. ACL, 2009.

[9] V. Zhechev and A. Way. Automatic generation of parallel treebanks. In *Proceedings of COLING 2008*, pages 1105–1112. ACL, 2008.