LinES: An English-Swedish Parallel Treebank

Lars Ahrenberg
NLPLab, Human-Centered Systems
Department of Computer and Information Science
Linköpings universitet
lah@ida.liu.se

Abstract

This paper presents an English-Swedish Parallel Treebank, LinES, that is currently under development. LinES is intended as a resource for the study of variation in translation of common syntactic constructions from English to Swedish. For this reason, annotation in LinES is syntactically oriented, multi-level, complete and manually reviewed according to guidelines. Another aim of LinES is to support queries made in terms of types of translation shifts.

1 Introduction

The empirical turn in computational linguistics has spurred the development of ever new types of basic linguistic resources. Treebanks are now regarded as a necessary basic resource (Nivre et al, 2005) and many of the parallel corpora that were created in the nineties are being developed into parallel treebanks. A parallel treebank extends the usability of a parallel corpus in several ways:

- The application of syntactic annotation schemes can be tested on several languages and enables multi-lingual evaluation and/or training of parsers.
- With access to syntactic relations and alignments we can provide much more fine-grained characterizations of structural correspondences and automatically identify and count such correspondences in the corpus.
- We can investigate the distribution of different kinds of shifts in different sub-corpora and characterize the translation strategy used in terms of these distributions.

In this paper the focus is mainly on the second aspect, i.e., on identifying translation correspondences of various kinds and presenting them to the user. When two segments correspond under translation but differ in structure or meaning, we talk of a translation shift (Catford, 1965). Translation shifts are common in translation even for languages that are closely related and may occur for various reasons. This paper has its focus on structural shifts, i.e., on changes in syntactic properties and relations.

Translation shifts have been studied mainly by translation scholars but is also of relevance to machine translation, as the occurrence of translation shifts is what makes translation difficult. While not all types of translation shifts need to be handled by a machine translation system at least the ones that are due to differences in grammar must be, and, generally speaking, the more of the others that can be handled, whether motivated by style or translator preferences, the better the system.

2 LinES

LinES, Linköping English-Swedish Parallel Treebank, is created on the basis of LTC, The Linköping Translation Corpus (Merkel, 1999). The selection of sentences from the sources are somewhat arbitrary. It has been assumed that whatever selection is made, as long as it is random, will provide typical examples of the usage of function words and grammatical

Eng. *Did you see the elephants ?*
Swe. *Såg ni elefanterna ?*
Links: (0,0,-1,-1,5)#(1,1,1,1,5)#(2,2,0,0,5)
#(3,4,2,2,5)#(5,5,3,3,5)

Figure 1: Encoding of word alignments in short format.

constructions and their translation.

2.1 Sub-corpora

The current version of LinES has two sub-corpora, Access, that includes sentences from MS Access online Help texts, and Bellow, with sentences taken from the novel *Jerusalem and Back* written by Saul Bellow. Each sub-corpus contains 600 sentence pairs that have been parsed and aligned at the word level. The goal is to include 1-2 more genres with different texts from each genre and about 1000 sentence pairs from each text.

A sub-corpus of LinES consists of three files: a source file, a target file, and a link file. Source and target files of LinES are XML-formatted monolingual files. These files are structured in terms of segments and words. Segments are demarcated by <s>-tags and words by <w>-tags.

A word normally corresponds to an orthographic word of the source text. However, punctuation marks and clitics are treated as separate words, and a restricted set of multi-word units, such as *of course, each other* are treated as single words.

Each segment has a unique identifier, its s-id. Corresponding source and target segments are assigned identical s-ids. Similarly, each word has a unique identifier, its word-id. In addition, each word has an identifier that states its relative position in the segment.

There are two formats for link files: an XML-format and a short format, where a correspondence is identified by five numbers. The first two numbers identify a word sequence from the source segment, and the next two numbers a word sequence from the target segment. (0,0) is the index for the first word. The pair (-1,-1) is used to represent a null alignment. The fifth number classifies the link as independent or as part of a discontinous alignment. An example of this encoding is shown in Figure 1.

2.2 Linguistic annotation

Words carry a number of attributes for linguistic annotation. The most important of these attributes are `base` for the word stem, `pos` for the part-of-speech, `msd` for morpho-syntactic properties, `func` for dependency relation with respect to a head word, and `fa` for the position of the head word.

Base forms are identical to one of the actually occurring forms of a word. Thus, the base form generally is not a proper lemma, as words of different parts of speech, and words of the same parts of speech with different inflections, may have the same base form.

A common set of parts of speech and morphological properties are used for both languages. While all part-of-speech categories apply to both languages, some morphological properties are used only for one of them. For instance, participial forms are subcategorized differently in English and Swedish, and only Swedish nouns are sub-categorized for definiteness.

The syntactic annotation in LinES is based on dependency relations. Each segment is assumed to have a single head token and all other tokens, except punctuation marks, are direct or indirect dependents of the head. The analysis is projective, i.e., no discontinuous phrases are allowed. This makes conversion to flat phrase structure representations simple. Dependency analysis has an advantage for parallel treebanks in that phrase alignment to a large extent is given for free from the word alignment.

For parsing, the Machinese Syntax parsers for English and Swedish from Connexor Oy, have been used[1]. These parsers supply initial values for base form, part-of-speech and morphological categorization. However, the annotation in LinES differ in several respects from the parser output. First, postprocessors that convert annotations and add morphosyntactic information not provided by the parsers are applied. Some function words have also been given different parts-of-speech in LinES.

The dependency functions used in LinES also differ from those of the parsers. The main difference is that they are structure-oriented. In particular, many functions with a primarily semantic flavour that the parses use are encoded as adverbials or modifiers,

[1] See `http://www.connexor.com/`

while others have been added. For instance, LinES distinguishes both prepositional objects and particles from adverbials and employs some additional functions not used by the parsers, such as vocative.

2.3 Alignment

Sentence alignment in LinES is taken over from LTC, while the guidelines for word alignment are slightly different. All of the alignments are manually reviewed, using the interactive word alignment system I*Link (Merkel et al, 2003).

The basic rule for alignment in LinES is the same as the one used in many other projects, namely "Align as short segments as possible, and as long segments as necessary". This guideline means that if we cannot find a good link for a word that we are looking at, we try to find a segment that includes that word that has a better correspondent. However, if the argument can go either way, we prefer many small links to few large ones. For example, a correspondence such as *the house* ∼ *huset* is aligned (0,0,-1,-1)#(1,1,0,0) rather than (0,1,0,0). Thus, so called level shifts (Catford, 1965) are normally encoded with the aid of null links in LinES.

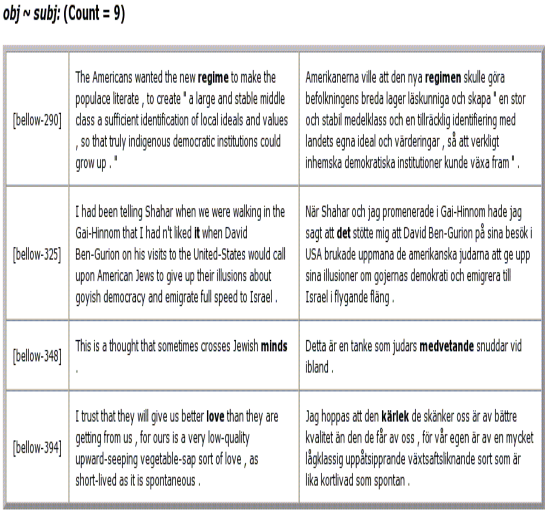
3 Querying LinES

A word-aligned parallel corpus can easily be queried for word correspondences, using whatever linguistic information is associated with the words. A parallel treebank can in addition be queried for functional information and, in principle, arbitrary subtrees and their correspondences.

The query interface for LinES is in development. The current web-based interface supports link-based search, while tree-based search is still in the pipeline.

3.1 Link-based queries

A link-based query can specify constraints on segments that have been aligned as a pair. In the simplest case the query specifies constraints on a single node of the dependency tree. In this case LinES supports any combination of constraints on base forms, parts-of-speech, morphological information, and dependency relation. Constraints can also be placed on the number of nodes. Moreover, constraints can be specified only for one of the languages, or for both of them. An example is shown in Figure 2.



obj ~ subj: **(Count = 9)**

[bellow-290]	The Americans wanted the new **regime** to make the populace literate , to create " a large and stable middle class a sufficient identification of local ideals and values , so that truly indigenous democratic institutions could grow up . "	Amerikanerna ville att den nya **regimen** skulle göra befolkningens breda lager läskunniga och skapa " en stor och stabil medelklass och en tillräcklig identifiering med landets egna ideal och värderingar , så att verkligt inhemska demokratiska institutioner kunde växa fram " .
[bellow-325]	I had been telling Shahar when we were walking in the Gai-Hinnom that I had n't liked **it** when David Ben-Gurion on his visits to the United-States would call upon American Jews to give up their illusions about goyish democracy and emigrate full speed to Israel .	När Shahar och jag promenerade i Gai-Hinnom hade jag sagt att **det** stötte mig att David Ben-Gurion på sina besök i USA brukade uppmana de amerikanska judarna att ge upp sina illusioner om gojernas demokrati och emigrera till Israel i flygande fläng .
[bellow-348]	This is a thought that sometimes crosses Jewish **minds** .	Detta är en tanke som judars **medvetande** snuddar vid ibland .
[bellow-394]	I trust that they will give us better **love** than they are getting from us , for ours is a very low-quality upward-seeping vegetable-sap sort of love , as short-lived as it is spontaneous .	Jag hoppas att den **kärlek** de skänker oss är av bättre kvalitet än den de får av oss , för vår egen är av en mycket lågklassig uppåtsipprande växtsaftsliknande sort som är lika kortlivad som spontan .

Figure 2: Output from LinES. The query concerns nodes with an object function in the source text corresponding to subjects in the target text.

3.2 Subtree search

In principle any subtree of a full dependency tree could be the object of an alignment relation. Moreover, if we wish to explain the occurrence of a certain structural shift, the relevant information may be located anywhere in the tree and even outside the tree. While it would be desirable to have a rich language for specifying tree queries, such as that used with Tgrep2 (Rhode, 2004), we do not initially aim for handling arbitrary combinations of constraints, but want to handle queries that classify a correspondence in terms of types of shifts such as deletion, addition, convergence, head switch, and so on.

We restrict consideration to subtrees that form a connected part of a full tree with a single node as its head and zero or more dependent nodes. If the head node has no dependents, the subtree and the node are identical.

A subtree is *inclusive* if it contains all (direct and indirect) dependent nodes of its head node. It is *unilevel* if its longest branch has length one, and is complete wrt to this depth if it contains all direct dependents. In addition to these two types of subtrees, queries for single branches and their correspondents need to be supported.

A subtree and its image are *isomorphic* if (i) they have the same number of nodes, (ii) the same number of branches, and (iii) the n:th branch of the image is an isomorphic image of the n:th branch of the given subtree, where n identifies left-to-right order. For a branch to be an isomorphic image of another branch we require that its m:th subtree of depth 1 is the image of the m:th branch of the other one.

Even if a subtree and its image are isomorphic, they are not necessarily free of shifts. This is so, because the notion of isomorphism so far defined does not take associated linguistic information into account. We believe that a simple formal solution is hard to find in spite of the fact that our categories are uniform. Thus, we need to treat correspondence in annotations notionally. Starting from a simple formal notion of regular correspondence for subtrees and their images, we may consider extending it by adding explicit equivalence relations that express normal relations when translating from English to Swedish.

4 Related work

Several projects for the creation of parallel treebanks have recently been launched. The FuSe project (Cyrus, 2006) annotates parts of the English and German sections of the Europarl corpus with regard to predicates and their arguments. LinES is different from FuSE in that it aims for complete alignments of segment pairs and (semi-)automatic derivation of shifts.

The CroCo-project (Hansen-Schirra et al, 2006) also works with German and English but has a larger scope. Complex queries based on the annotation for many types of shifts can be formulated, though so far only with detailed knowledge of the XML-format and the details of the annotation.

The SMULTRON corpus (Volk et al, 2006; Samuelsson and Volk, 2006) includes data from three languages (English, German, and Swedish). The annotation is based on phrase structure analyses. This project is primarily oriented towards machine translation and the recognition of translation equivalents that can “serve as translations outside the current sentence context” (Samuelsson and Volk, 2006). For this reason, phrase alignment of a sentence pair need not be complete and, contrary to LinES, the alignment of non-equivalent phrases are avoided rather than sought for.

References

J. C. Catford 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*, London, Oxford University Press

Lea Cyrus 2006. Building a resource for studying translation shifts. Proceedings of LREC 2006, Genoa, May 24–26, 2006 1240-1245.

Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 2006. Multi-dimensional Annotation and Alignment in an English-German Translation Corpus. Proceedings of the EACL Workshop on Multi-dimensional Markup in Natural Language Processing (NLPXML-2006), Trento, Italien, 4. April 2006, 35-42.

Magnus Merkel 1999. Understanding and enhancing translation by parallel text processing. *Linköping Studies in Science and Technology, Dissertation No. 607.* Linköping, 1999.

Magnus Merkel, Michael Petterstedt, and Lars Ahrenberg. Interactive Word Alignment for Corpus Linguistics. Proceedings of Corpus Linguistics 28-31 March, 2003, Lancaster. UK.

Joakim Nivre, Koenraad de Smedt, and Martin Volk 2005. Treebanking in Northern Europe: A White Paper. Holmboe, Henrik (ed.) Nordisk Sprogteknologi 2004: Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004, 97-112. København: Museum Tusculanums Forlag.

Yvonne Samuelsson and Martin Volk. 2006. Phrase Alignment in Parallel Treebanks Jan Hajiĉ and Joakim Nivre (eds.) *Proceedings of the Treebank in Linguistic Theory Workshop*, Prague, Czech Republic, December 2006. 91-102.

Douglas T. L. Rhode 2004. TGrep2 - the next-generation search engine for parse trees. `http://tedlab.mit.edu/~dr/Tgrep2/`.

Martin Volk, Sofia Gustafson-Capková, Joakim Lundborg, Torsten Marek, Yvonne Samuelsson, and F. Tidström. 2006. XML-Based Phrase Alignment in Parallel Treebanks. *Proceedings of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*, Trento, Italy, April 2006.