Alignment Guidelines for SMULTRON

Yvonne Samuelsson, Martin Volk, Sofia Gustafson-Capková, Elisabet Jönsson Steiner
Stockholm University, Department of Linguistics
University of Zürich, Institute of Computational Linguistics

Version 2.1, August 9, 2010

Contents

1 Introduction

Creating alignment in parallel treebanks is to annotate which part of a sentence in one language is equivalent to which part of a corresponding (= translated) sentence in another language. We approach this goal by drawing alignment lines between sentences, phrases and words over parallel trees (which represent the syntactic structure of the underlying sentences) with the help of the Stockholm TreeAligner. One simple example is shown as a tree structure in Figure 1.

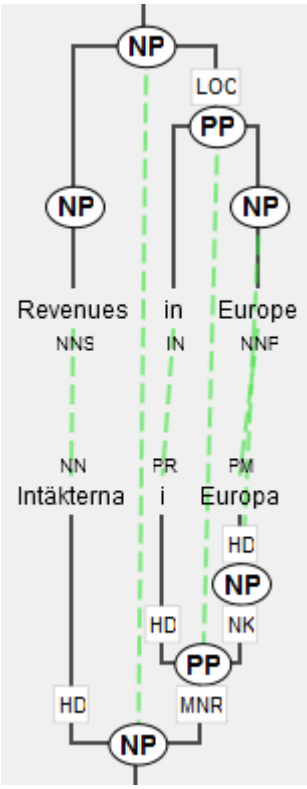

Figure 1
Example of word and phrase alignment.

There are basically two types of alignment, phrase alignment and word alignment. Phrases are nonterminal nodes in the tree structures, and can therefore contain everything from one word to a whole sentence. Word alignment is drawn between terminal nodes. In the following, if we refer to a specific kind of node this will be spelled out (e.g. *phrasal node* or *word node*), whereas the term *node* on its own will be used to refer to any node in general. Alignment is either word-to-word or phrase-to-phrase.

We want to align as many phrases and words as possible. The goal is NOT to show parallelism in the translations per se, but to show translation equivalence. Phrases shall only be aligned if the tokens that they span represent the same meaning. They represent the same

meaning if they could serve as translation units in an Example-based Machine Translation system. In other words, they represent the same meaning if they can be used as translation units outside the current sentence context. Note that the grammatical forms of the words or phrases need not fit in other contexts, but the meaning has to fit.

Throughout these guidelines examples are given in English, German and Swedish. In the examples, text is sometimes given in (parenthesis) to denote the context of the actual example, whereas [brackets]NP denote a phrase**.** Annotation follows the Penn Treebank guidelines for English, the TIGER Annotation guidelines for German and the SWE-TIGER Annotation guidelines for Swedish. The examples may, however, show simplified versions of the structures, to focus on the particular problem being discussed.

2 Alignment

As stated, we align as many words and phrases as possible, that represent the same meaning. We only align words and phrases that can serve as translation units outside the current sentence context. This means that phrases like in example (1) should NOT be aligned, even though they can be seen as equivalent in this particular translation.

(1) **SV:** *(han var) [ute på världshaven]PP*

EN: *(he was sailing) [all over the world]PP*

In isolation the Swedish phrase *ute på världshaven*, which literally means *out on the Seven Seas*, is not equivalent to *all over the world* since the information about sailing or sea is not contained in the English phrase.

It also means that alignment showing more precise information is better than alignment showing more general information. In example (2) this means that the English NP *nothing* should be aligned to the Swedish NP *ingenting* rather than the coordinated NP (CNP) *noll och ingenting*.

(2) **SV:** *(ha blivit till av) [noll och [ingenting]NP]CNP*

EN: *(have come from) [nothing]NP*

2.1 Exact alignment

If the phrases represent exactly the same meaning, like in examples (3), (4) and (5), then they are aligned with a green line (= exact alignment).

(3) **DE:** *[aus null und nichts]PP*
EN: *[from nothing]PP*

(4) **DE:** *[von der Schule]PP*
EN: *[from school]PP*

(5) **DE:** *[sich über Roboter unterhalten]VP*
EN: *[discussing robots]VP*

Even if the word forms differ in grammatical properties (e.g. definiteness vs. indefiniteness; singular vs. plural), they can be aligned with exact alignment if their meanings overlap to a large extent.

2.2 Fuzzy alignment

If the nodes represent approximately the same meaning, then they are aligned with a red line (= fuzzy alignment). This applies to example (6), where the PP-nodes get fuzzy alignment because of the pronoun *her*.

(6) **DE:** *[auf dem Heimweg von der Schule]PP*
EN: *[on her way home from school]PP*

Example (7) requires fuzzy phrasal node alignment because of the hyperonymy relation between *Maschine* and *hardware*.

(7) **DE:** *[mehr als eine Maschine]AP*
EN: *[more than a piece of hardware]ADJP*

Example (8) requires fuzzy alignment between the PP-phrases because of the additional information on the houses in German.

(8) **DE:** *[am Ende eines ausgedehnten Viertels mit Einfamilienhäusern]PP*
EN: *[on the outskirts of a sprawling suburb]PP*

Equivalent names, as in example (9), are assigned exact alignment even if they are spelled slightly differently. If the name has been exchanged, as in example (10), fuzzy alignment is required.

(9) **DE:** *[Sofie]NP*
EN: *[Sophie]NP*

(10) **DE:** *[Jorunn]NP*
EN: *[Joanna]NP*

A word can be aligned to an empty word if that empty word stands for an omitted (rather than moved) word node. Alignments to empty words are always fuzzy. In Figure 2 the German sentence contains the conjunction *ob* whereas an empty node (–) in the English sentence indicates that the conjunction has been left out. In this and equivalent cases the empty word should be aligned.[1] Note that the phrasal nodes that are closest to the word level and that contain the empty word alignment also get fuzzy alignment. This can also be seen in Figure 2 where the German S-node and the English SBAR-node are aligned with fuzzy alignment.

[1] This has not been consistently done and needs to be updated in the SMULTRON treebank.

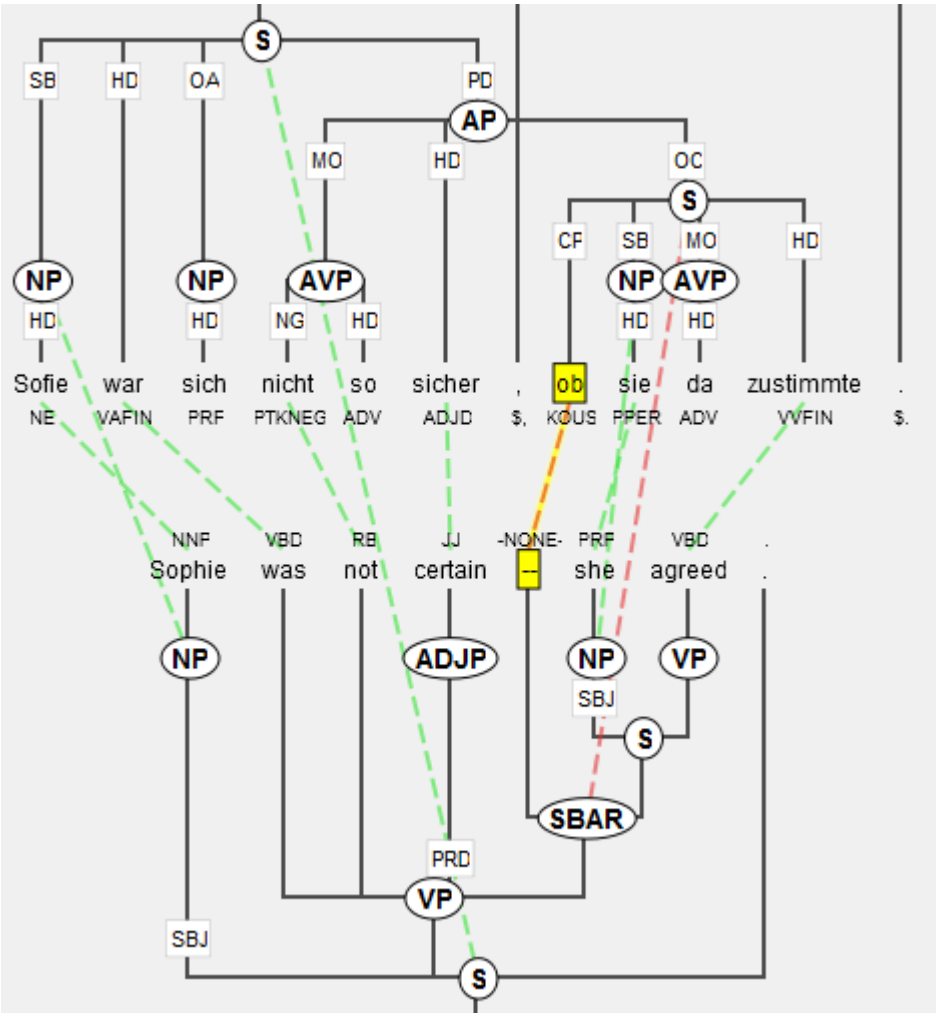

Figure 2

Example of empty word alignment where both word node alignment and phrasal node alignment closest to word level are fuzzy; Example of exact node alignment on a higher level despite fuzzy alignment on a lower level.

It is possible that phrases are aligned with exact alignments although some parts of them are aligned with fuzzy alignments. In Figure 2 above, although there is fuzzy alignment on lower levels, in this case due to an empty word in one language, there is exact alignment on higher node level, since the sentences match as good translations as a whole.

2.3 Multiple alignment

Nodes from one tree can be aligned to one or more nodes in a corresponding tree. If one node in L1 is aligned to several nodes in L2, these links are all of the same type, either exact or fuzzy.

2.3.1 Multiple sentence alignment

We allow m:n (i.e. one-to-many and many-to-many) sentence alignments. Thus, one or more nodes from one sentence can be aligned to one or more nodes from sentences across different trees. This is exemplified in Figure 3.

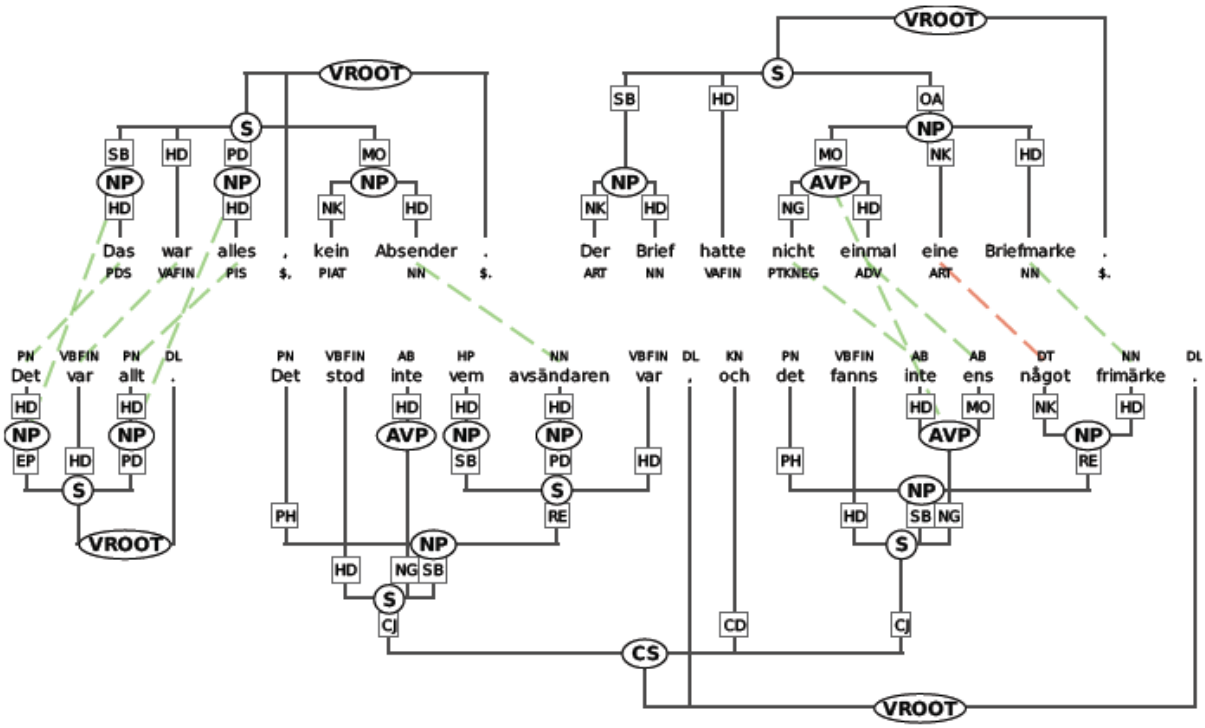

Figure 3
Example of m:n sentence alignment.

2.3.2 Multiple phrase alignment

We strongly discourage many-to-many alignments of nodes on sub-sentential levels. We do allow 1:n (one-to-many) alignments of phrasal nodes, even though it is not very common. Thus, a phrasal node from one tree can be aligned to one or more phrasal nodes in the corresponding tree. One example is in Figure 4, where the adverb could be in the PP or outside it. In this case it is inside the Swedish PP but outside the English PP. Therefore both the ADVP node *home* and the PP node *from school* should be aligned to the Swedish PP node *hem från skolan*. In Figure 5 the negation is moved to a position before the verb, outside of the NP, for Swedish, while it is inside the NP in English.

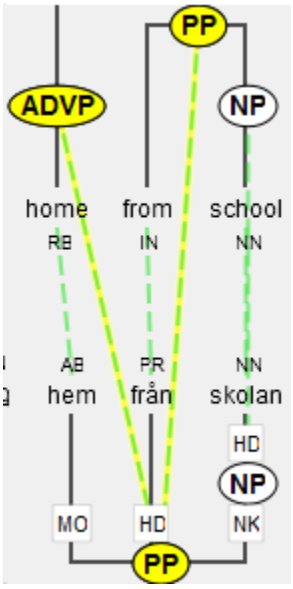

Figure 4
Example of 1:n exact phrasal node alignment.

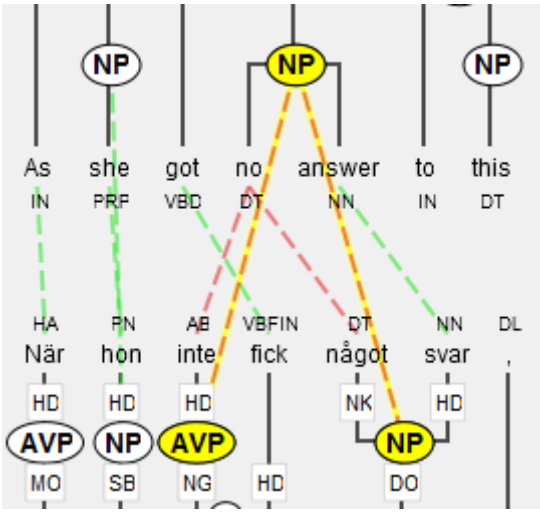

Figure 5
Example of 1:n fuzzy phrasal node alignment.

If a phrasal node has a unary mother node (e.g. a multi-part name, PN, used as a subject, NP, in the DE and SV annotations), then the phrasal node that is closest to the word level is to be chosen for alignment (in the example in Figure 6 the PN-node).

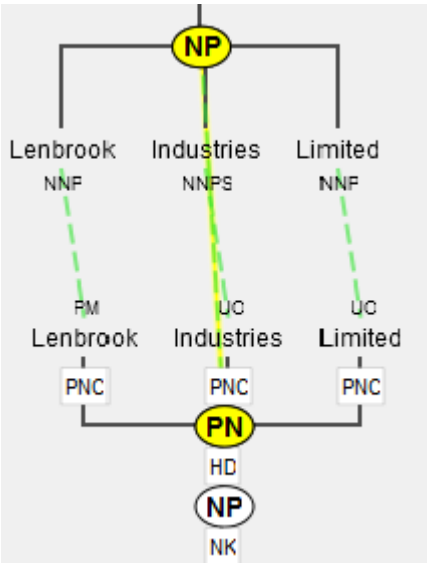

Figure 6
Example of alignment with a unary mother node.

2.3.3 Multiple word alignment

We strongly discourage many-to-many alignment of nodes on word level but we do allow 1:n word alignment. It is however important not to force word alignment and create 1:n links. In those cases alignment can usually be drawn at the phrasal node level instead, to show equivalence. In example (11) *looked* should be aligned to *sah* and *an***.**

(11) **DE:** *[Sie sah noch einmal den Briefumschlag an]S*

EN: *[She looked at the envelope again.]S*

Note that German *an* is not aligned to English *at*. Separated verb prefixes in German shall not be aligned to prepositions in the other language.[2] See separate paragraph in section 2.5.2 for alignment of prepositions and particles. In Figure 7 there are three instances of one Swedish word being aligned to two English words.

[2] This has not been consistently done and needs to be updated in the SMULTRON treebank.

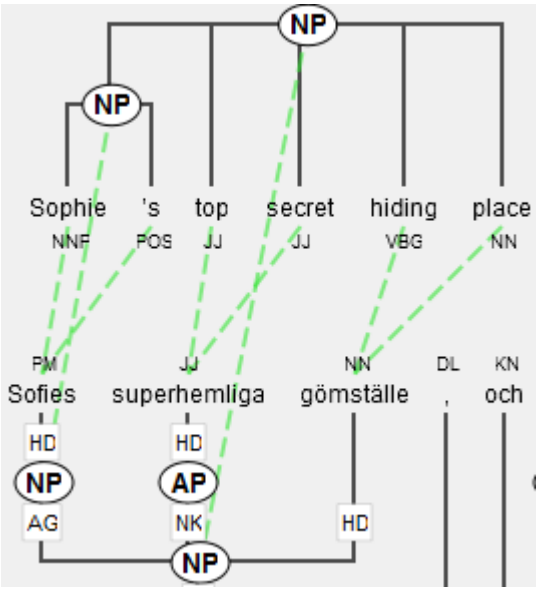

Figure 7
Example of multiple word alignments.

The English genitive marker *'s* is aligned as if it were a part of the preceding word. As we can see in Figure 7 above this means alignment from the Swedish *Sofies* to both *Sophie* and *'s*.

In (12) we cannot draw word alignment between *reser sig* and *gets up*, since many-to-many alignment would be required. In many cases this can be remedied with alignment on a higher level, otherwise it is simply missing alignment.

(12) **SV:** *(Efter ett tag) reser sig (mamman)*

EN: *(After a while Mom) gets up*

2.4 No alignment

The multiple alignment option is only used if a phrase/word from one tree corresponds in meaning to more than one phrase/word. It shall not be used if a phrase from one tree is realized twice in the corresponding tree, like in example (13). The German pronoun *sie* is realized as *they* twice in the English sentence. It will only be aligned to one of them. The annotator may choose the one that is most appropriate.

(13) **DE:** *[Beim Supermarkt hatten [sie]NP sich getrennt]S*

EN: *[When [they]NP got to the supermarket [they]NP went their separate ways .]S*

Pronouns should not be aligned to full noun phrases, i.e. no alignment between *Der Brief* and *it* in example (14).

(14) **DE:** *[[Der Brief]NP hatte nicht einmal eine Briefmarke]S*

EN: *[There was no stamp on [it]NP either .]S*

The English annotation includes the punctuation in the tree structure, which the German (and Swedish) annotation does not. This means that a sentence can be aligned regardless of wether punctuation is present or not. Punctuation marks are also not aligned at word level.

2.5 Problem areas

There are of course many problematic cases, some due to the translator's freedom. This can be seen both on the word level and on the sentence level. In these cases it is important to remember the main goal, that the aligned phrases should be equivalent outside of the sentence context.

2.5.1 More information in one language

There is always a question about how much extra information can be contained in a phrase in one language before it is too much for considering it equivalent in meaning. Because the goal is to show translation equivalence, we take a rather restrictive approach with regards to exact alignment. One additional adjective in one language may be acceptable when aligning noun phrases, or punctuation symbols (which are included in the English sentence, but not the Swedish or German).

However we do not want to align a full Swedish or German sentence S (which include the subject) with an English VP (which does not include the subject). Example (15) shows two sentences, the Swedish containing two coordinated sentences (one without subject) and the English containing two coordinated verb phrases. The Swedish CS is aligned to the English S and the second Swedish S, *skrek*, is aligned to the second English inner VP, *screamed*. The first Swedish S, *Han skrattade*, however cannot be aligned to the first English VP, *laughed*, since the Swedish node contains the subject. This problem is quite frequent in our treebank, due to the differences in annotation schemas between the languages.

(15) **SV:** *[[Han skrattade]S och [skrek]S]CS*

EN: *[He [[laughed]VP and [screamed]VP]VP .]S*

In example (16) the Swedish NP contains a relative clause, *som...*, and therefore it cannot be aligned to the English NP. Thus the NP's can only be aligned on word level (*hair* to *hår* and *straight* to *raka*).

(16) **SV:** *[det [raka]AP håret [som...]S]NP*

EN: *[her straight hair]NP*

English reflexive pronouns distinguish between gender, which is not the case for e.g. Swedish or German. This means that there is extra information in the English *herself/himself/itself* as compared to the Swedish *sig* or German *sich*. Basically we would like to have one way alignment, saying that *herself* is always translated by *sig*, while not the other way around. Since it is not possible to show that alignments are exact in only one direction but not in the other direction, they need to be linked through fuzzy alignment.[3]

2.5.2 Different ways of expressing the same thing

As said before grammatical forms need not fit in other contexts for there to be alignment. Some issues however go beyond grammatical forms like case and number. If the sentence in one language is active and the other is passive, pronouns in these sentences should get fuzzy alignment as in Figure 8.[4] In the example the German active sentence has the personal pronoun *sie* as an object form, whereas the English passive sentence has the personal pronoun *her* as a subject form. We do not want to make subject forms and object forms equal, since the translation will be wrong outside the current context.

[3] In some of the SMULTRON treebanks (e.g. Sophies World DE-EN) such cases have erroneously been assigned exact alignment. This should be corrected to fuzzy alignment.

[4] This has not been consistently done and needs to be updated in the SMULTRON treebank.

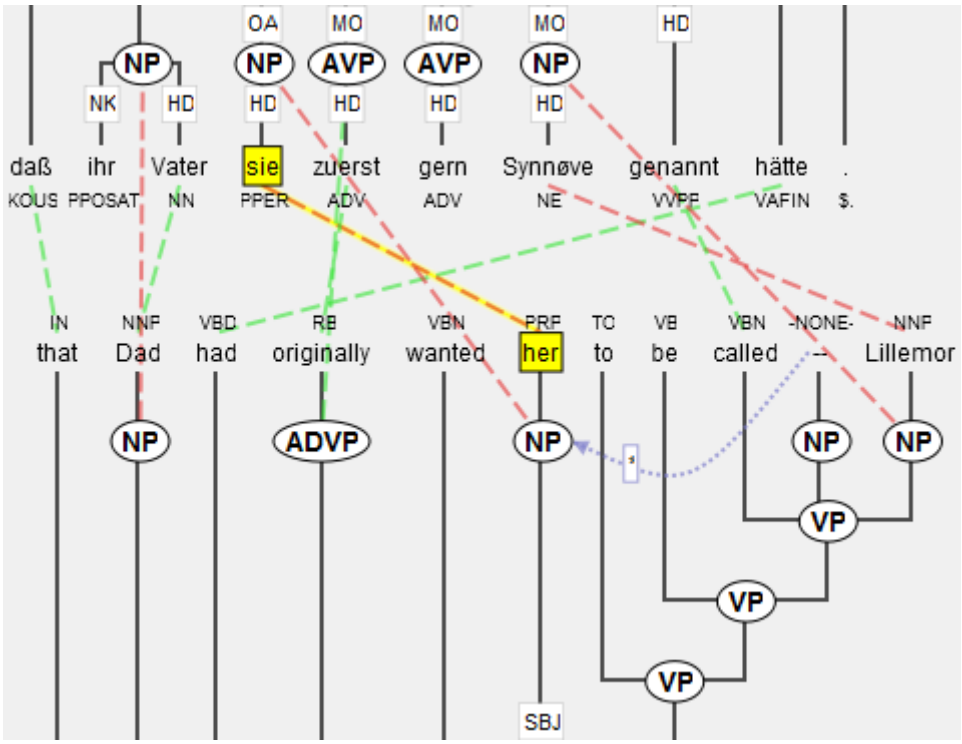

Figure 8
Example of active sentence in one language vs. passive sentence in the other, i.e. object form vs. subject form of the pronouns, requiring fuzzy alignment.

Prepositions should be aligned with exact alignment if the structure of the prepositional phrase is the same in both languages, like in example (17) *vid* - *at*. In some cases 1:n alignment between a preposition in one language and a preposition plus a verb particle in the other language is required, like in example (18). There the Swedish verb particle *in* together with the preposition *på* needs to be aligned with the English preposition *into*.

(17) **SV:** *vid (den här tiden på året)*
EN: *at (this time of year)*

(18) **SV:** *(svängde) in [på [Klövervägen]NP]PP*
EN: *(turned) [into [Clover Close]NP]PP*

Verb particles are aligned differently from prepositions. As the example in Figure 9 shows, in cases where there is a verb+particle (*tryck in*) in L1 represented by only a verb in L2 (*press*), then both verb and particle in L1 are aligned to the verb in L2. Compare Figure 10, where the preposition *på* is not aligned to the verb.

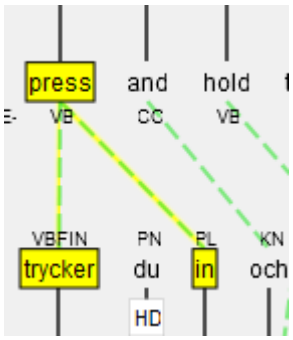

Figure 9
Example of particle being aligned in the case of verb vs. verb+particle.

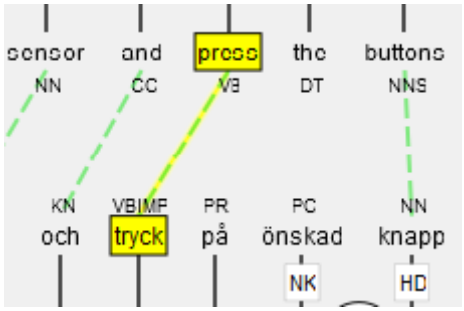

Figure 10
Example of preposition not being aligned in the case of verb vs. verb+preposition.

Note that, as mentioned in section 2.3.3 above, prepositions should not be aligned to separated verb prefixes in German.

In the case of German contractions of pronouns and articles this should be aligned with exact alignment to both preposition and article in the other language. An example is given in (19), where *am* is aligned to *on* and *the*.

(19) **DE:** *[am [Ende]NP]PP*

EN: *[on [the outskirts]NP]PP*

There are cases where an infinite verb has the infinitive marker in English, e.g. *to think*, but this can been left out in the other language e.g. Swedish *tänka*. In such cases the infinitive marker should not be aligned to the verb as in (20).

(20) **EN:** *(She tried) to think (extra hard)*

SV: *(Hon försökte) tänka (extra hårt)*

When a verb needs a reflexive pronoun in both languages, e.g. English *repeat itself* and German *wiederholt sich* in (21), the verbs are aligned to each other *repeat* - *wiederholt* and the reflexives to each other *itself* - *sich*. If the reflexive is only necessary in one of the languages, e.g. in German *rasiert sich* but not in English *shaves* in (22), both the verb and the reflexive are aligned to the verb (without reflexive) to show this difference in necessity.

(21) **EN:** *(This rapturous performance may) repeat itself*

DE: *(Vielleicht) wiederholt sich (diese wüste Szene)*

(22) **EN:** *(Every day Dad) shaves*

DE: *(Jeden Tag) rasiert (er) sich*

If, L1 has a finite verb, e.g. German *kommt*, while L2 has an infinite verb with an auxiliary verb, e.g. English *will come,* both L2 verbs should be aligned to the L1 verb as in Figure 11. If both languages contain an auxiliary (plain auxiliary or modal auxiliary) and an infinite verb, e.g. English *can experience* and German *kann erleben*, then the auxiliaries are aligned to each other, and the infinite verbs to each other as in Figure 12. These rules of course only apply if the main verb has the same meaning (i.e. should be aligned at all). Note that different auxiliaries that have the same function in the sentences should be aligned. However, modal auxiliaries and plain auxiliaries should not be aligned.

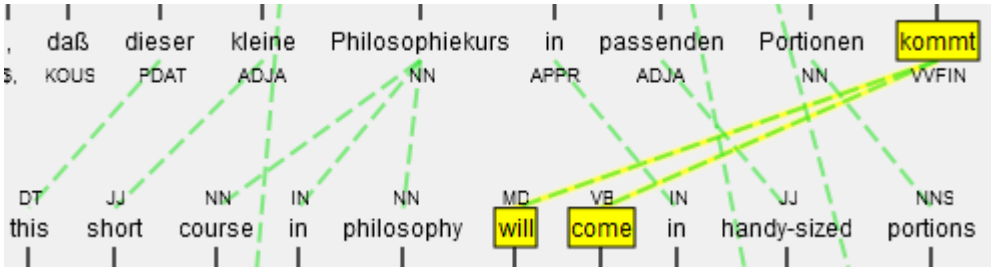

Figure 11
Example of L1 finite verb vs. L2 auxiliary + infinite verb alignment.

Figure 12
Example of alignment with auxiliary + verb in both languages.

There are cases where an L2 translation contains both the L1 equivalent expression as well as an L2 translation of the same. In the example in Figure 13 the Swedish sentence contains both the English expression *NO DISC* as well as the Swedish translation *INGEN SKIVA*. In such cases word alignment should be between the actual translation and the English expression.

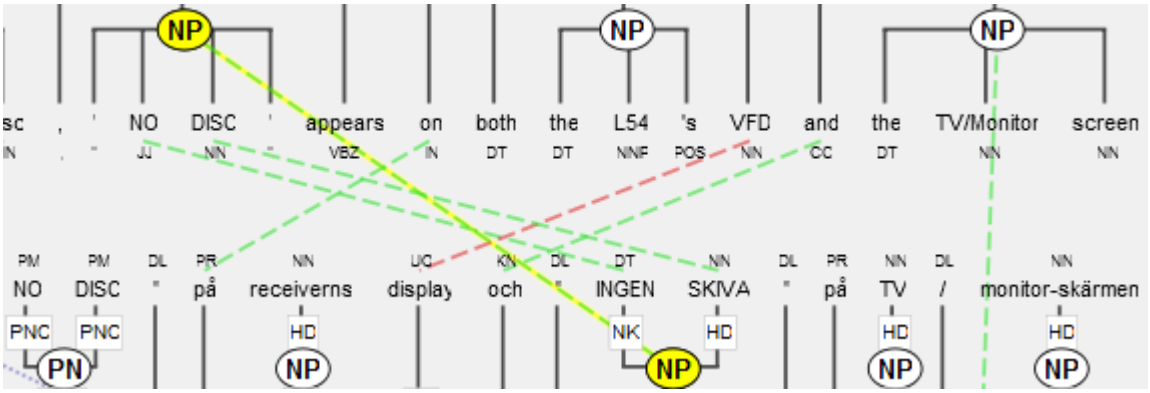

Figure 13
Example of word alignment with preference of actual translated expression.

Note that in cases where the translation only gives the same expression as in the L1 text, these should be aligned.

A language specific difference worth mentioning, is the Swedish expression *för att* which can either be equivalent to English *because* or *to/in order to*. Figure 14 shows an example of the expression *för att* as a conjunction meaning *because*. In such cases *för* and *att* are both labelled as SN (subjunctions) and should be aligned to *because*. However, there are other cases where *för att* has the meaning of *in order to* but is usually equivalent to English *to* as in Figure 15. In this context *för* is labelled as PR (preposition) and *att* as IE (infinitive marker) and only *att* should be aligned to English *to.* This is the same case as German *um zu*, where only the infinitive marker should be aligned between English and German (*to-zu*). Obviously

both prepositions and infinitive markers should be aligned between Swedish and German (*för-um, att-zu*).

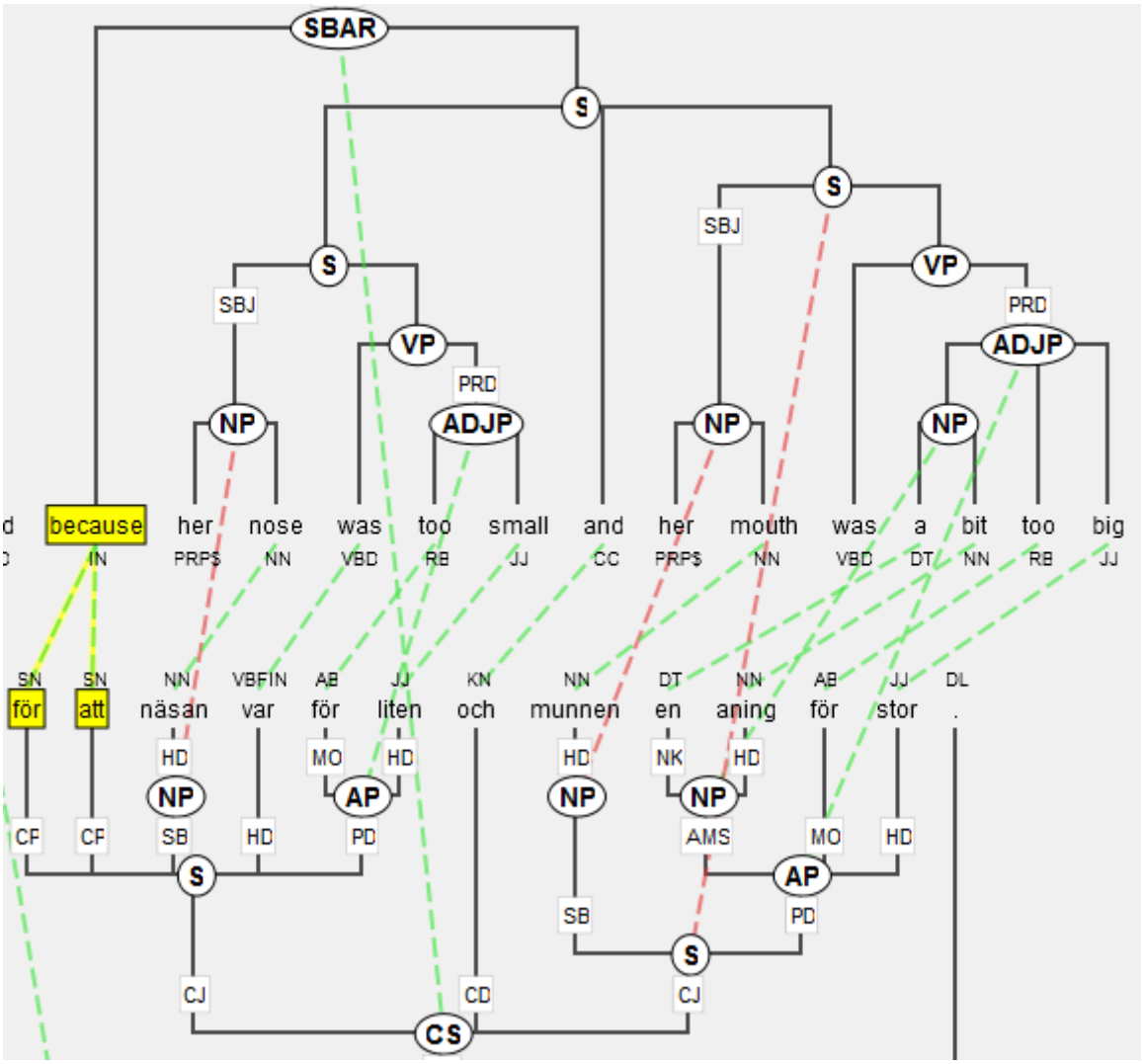

Figure 14
Example of alignment of *för att* vs. *because*.

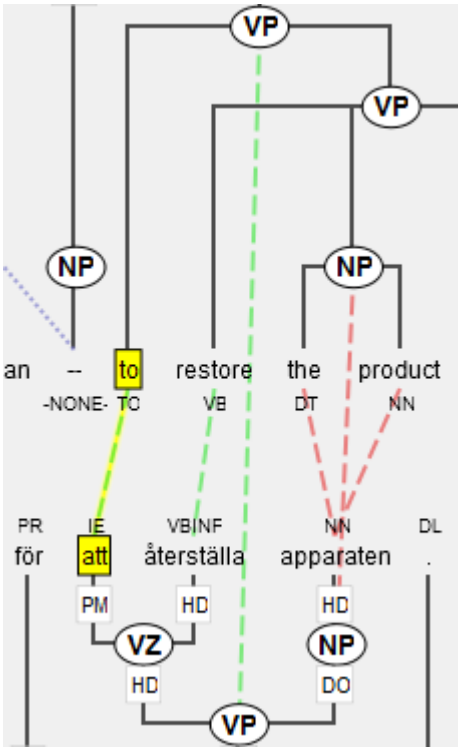

Figure 15
Example of alignment of *för att* vs. *to*.