Clause restructuring in English-Swedish translation

Lars Ahrenberg

Department of Computer and Information Science
Linköping University
E-mail: lars.ahrenberg@liu.se

Abstract

Medium rank clauses, such as participial and infinitival clauses, have been shown in earlier studies to be more frequent in English than in Swedish. Instead Swedish prefers complete, finite clauses. This constitutes a problem for English-Swedish machine translation. Here I report a study of such constructions using the LinES Parallel Treebank. I also show how the dependency annotation in LinES can be used to define clauses of different ranks.

1 Introduction

Clause structure is a major aspect of syntax and also one in which languages differ. Mastering clause structure means not only being able to produce grammatically well-formed clauses, but also being able to select the right type of clause in the right context. This is of special importance in translation, where source language norms may clash with target language norms. A good human translator would have developed a good sense of linguistic differences in this respect, but machine translation systems are often vulnerable to the influence of source language clause structure.

In statistical machine translation the type of restructuring that has been considered most is reordering. Several studies have shown improved performance when clause constituents are reordered to meet the norms of the target language e.g., [5], [6]. However, reordering is not the only relevant aspect of restructuring; additions and deletions of major constituents may be necessary or preferred as well as shifts of verbal morphology. In this paper our focus is on tenseless and subjectless constructions, which tend to be more common in English than in Swedish. The following are two examples where we compare a human translation to the translation suggested by Google Translate[1]:

[1]The first example is a variant of an authentic example from the LinES corpus, with translation by the author, the second example is taken from the Harry Potter section of LinES.

(1) EN: When creating a copy, she uses a very sharp point.
SE: När hon gör en kopia använder hon en mycket vass udd.
Gloss: When she creates a copy, she uses ...
Google: När du skapar en kopia, använder hon en mycket vass spets.
Gloss: When you create a copy, she uses ...

(2) EN: She leered at him, showing mossy teeth.
SE: Hon gav honom ett illvilligt leende, som avslöjade maskstungna tänder.
Gloss: She gave him a malicious smile that revealed mossy teeth.
Google: Hon sneglade på honom, visar mossiga tänder.
Gloss: She leered at him, shows mossy teeth.

For both examples Swedish prefers a finite clause construction with an overt subject or subject place-holder. This means generating both a tense and a subject that are congruent with the context. Google Translate manages to generate a tense, and, in sentence (1) also a subject, but does not succeed in enforcing the requirements on congruence. While other translations are possible here, literal translations using a Swedish participial form are not.

In linguistically oriented translation studies changes of this kind have been studied and classified by many authors, e.g., as category shifts [4], or transpositions [11]. Since they introduce material which is only implicit in the source, they may also be regarded as explicitations. Here I will call them shifts of clause rank, or simply rank shifts, after [7]. In this work Rune Ingo compares Finnish and Swedish on the one hand, and English and Swedish on the other. One claim of his is that participial and infinitival clauses are more common in Finnish and English than they are in Swedish, supporting the claim with percentages from different kinds of corpora. He also argues for the position that, normally, the translator should produce different types of clauses in the proportions that are suitable for the target language and the given text genre.

In this paper I treat Ingo's claim as an hypothesis to be tested for English source texts and Swedish translations, using the LinES English-Swedish parallel treebank [1] as test data. The hypothesis, then, is that the English half of the corpus contains more instances of participial and infinitival clauses than the Swedish side, and more specifically, that we will find a significant number of cases, where such clauses have been translated by clauses of a higher rank, in particular finite clauses. I will also investigate to what extent there are differences among the different text types included in LinES.

This, in turn, raises the question whether different clause types can be recognized with high accuracy in a corpus where the syntactic annotation does not explicitly mark them. Thus, another aim of the paper is to provide definitions of various clause types using the annotation in the treebank. The study as a whole can be taken as a support for the view that syntactically annotated parallel corpora are useful for translation studies. While parallel corpora have been recognized as

primary sources of data in many areas including translation studies and translation training ([3], [9]) they are not usually annotated syntactically. However, the range of linguistic and translational phenomena one can study is very much dependent on the available corpus annotation.

The rest of the paper is organized as follows. In the following section I introduce Ingo's model of clause ranks [7]. Section 3 presents our data and the annotation used. Section 4 explains the method and the use of the annotation for the purpose of the study. Section 5 presents our findings, and Section 6, finally, states the conclusions.

2 Clause types and clause ranks

Ingo's model of clause ranks has six levels:[2]

major clause,	e.g. *they arrived in London; Kim plays the guitar*
minor clause,	e.g. *when they arrived in London; that Kim plays the guitar*
participial clause,	e.g. *arriving in London; (while) playing the guitar*
infinitival clause,	e.g. *(them) to arrive in London; (ask her) to play the guitar*
nominalization,	e.g. *their arrival in London; Kim's guitar play*
without predication,	e.g. *in London, they... ; on guitar, Kim*

The further down we go in this hierarchy, the more the constructions lack features of a complete clause. These features are, according to Ingo, (i) the presence of a subject, (ii) the presence of a tense marker, (iii) the marking of mode, (iv) the optional presence of a negation.

3 The data

The parallel treebank used for this study comprises four subcorpora as outlined in Table 1: on-line help texts for MS Access for Windows XP (Access), Europarl data, and excerpts from two novels. Each subcorpus used for the study has a size of roughly 600 sentence pairs. The syntactic annotation employs parts-of-speech, morphological properties, and dependency functions. Every sentence is assumed to have a unique head, marked by the function 'main', and all other tokens, except punctuation marks, are direct or indirect dependents of the head. Monolingual files are XML-formatted. An annotated segment pair is shown in Figure 1.

The dependency annotation employed in LinES is surface-oriented and projective, making it easy to convert into a phrase-structure tree. The monolingual files were first parsed by Connexor parsers for English and Swedish [10] but the actual

[2]The English terms are translations from the Swedish ones used by Ingo: *clause rank: satsgrad; major clause: huvudsats; minor clause: bisats; participial: particip; infinitive: infinitiv; nominalization: nominalisering; without predicate: predikatslös.*

```
<s id="s3">
<w .. relpos="1" base="noone" func="subj" fa="2" pos="PRON" msd="SG">Noone< /w>
<w .. relpos="2" base="be" func="main" fa="0" pos="V" msd="PRES">is< /w>
<w .. relpos="3" base="very" func="ad" fa="4" pos="ADV">very< /w>
<w .. relpos="4" base="patient" func="sc" fa="2" pos="A" msd="ABS">patient< /w>
<w .. relpos="5" base="." pos="FE" msd="Period">.< /w>
< /s>
```

Figure 1: Morphosyntactic annotation of an English sentence in LinES.

annotation employs a different set of values, and for some constructions, different analyses. All annotations, including dependencies and alignments have been manually reviewed.

Subcorpus	**Text type**	**Sentences**	**Src words**	**Trg words**	**Ratio**
Access	online help texts	595	10,451	8,898	0.85
Europarl	political debates	594	9,334	8,715	0.93
Bellow[3]	fiction	604	10,310	9,962	0.97
HarryP[4]	fiction	600	10,171	10,501	1.03
Sums:		2393	40,266	38,076	0.95

Table 1: Corpus overview showing text type and size.

The word alignment is based both on semantic and structural correspondence where many-to-many alignments (as usual) represent corresponding units that cannot be analysed into smaller (1-1, 1-n, or n-1) alignments. Alignment was performed interactively using the I*Link tool [8]. Word alignments are complete, i.e., a decision has been made for each token in the corpus if, and how, it corresponds to something in the other language. A word link is represented as a paired list of indices such as (4-5/1) which says that the 4th and 5th words of the source sentence have been linked to the first word of the target sentence. The alignment encoding for the sentence in Figure 1 and its Swedish translation is shown in Figure 2.

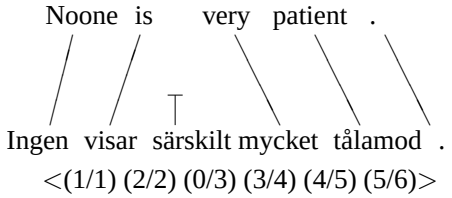

Figure 2: Swedish translation and alignment for the English sentence in Figure 1.

Null links are represented by the number 0. For example, (0/3) means that the third word of the Swedish sentence is judged to have no correspondent in the

English sentence.

4 Defining clause ranks

Our basic approach is to identify all clauses in the corpus, and then classify them in terms of a given clause rank. We restrict this process to clauses that are governed by verbs or participles, i.e., to words that have the part-of-speech annotation pos="V" or pos="PCP". Since each clause has a single governor we can identify corresponding pairs of clauses through the word alignment. If the translation is not a clause, but a phrase of some sort, we can still identify the image and its properties.

4.1 LinES annotation of clause elements

The notion of clause is underlying the LinES annotation, since a subset of the dependency relations are defined to apply only at clause level, i.e., to relate words to a verbal item. Similarly, other dependencies are restricted to noun phrases, while others, such as the coordination dependency can have almost any type of governor. The clause-level dependency relations used in LinES are listed in Table 2.

Distinctions between different clause types, however, are not part of the annotation scheme, and cannot be seen in the definitions of categories or dependency relations. In particular, Ingo's system of clause ranks formed no part in its development. It is therefore something of a test for the LinES scheme to model clause ranks within the scheme.

Label	Explanation
vch	Auxiliary verb or infinitive marker
advl	Adverbial
subm	Subjunction
subj	Subject
obj	Object (direct or indirect)
sc	Subject complement
oc	Object complement
prt	Particle
pobj	Oblique object
initm	Initiator e.g. an interjection
ad	General pre-modifier
cc	Coordinating conjunction or conjunct

Table 2: Clause-level dependency relations in LinES annotation. Clause-specific relations above the horizontal line.

4.2 Clause rank definitions

For the definition of clause ranks we need to consider the properties of the clause governors and the relations to their dependents. In some cases, such as the property of being tensed, the subtree dominated by a clause governor needs to be searched into sufficient depth. As in [10] auxiliaries, including the infinitive markers, are related to their governors via the dependency 'verbal chain element' (vch). Usually the tense marker will appear on the first element of the verbal chain, and the whole chain needs to be searched. Also the subject is a dependent of the first element of the verbal chain, rather than the main verb, when there are auxiliary verbs.

Clause rank	**Features**
major clause	+Tensed, +Subject, +Main
minor clause	+Tensed, +Subject or +Subjunction, -Main,
coordinated VP	+Tensed, -Subject, +ccVerb
participial clause	-Tensed, -Subject, +Participial, -Attributive
participial attribute	-Tensed, -Subject, +Participial, +Attributive
infinitival clause	-Tensed, -Subject, +Infinitive

Table 3: Clause rank feature analysis.

The main features that distinguish the different clause ranks are the following:

+Tensed, the presence of a tense marker on the first element of the verbal chain.

+Participial, the chain is -Tensed and the first element is a past or present participle.

+Infinitive, the chain is -Tensed and the first element is the infinitive marker or a verb in infinitive form.

+Subject, the presence of a word contracting the subject relation to the verbal chain. Imperative verbs usually don't have explicit subjects, but are assigned this feature by default.

+Subjunction, The presence of a subjunction or phrase having the 'subm' relation to the verbal chain.

+Main. the property of being the governor of an entire segment. The opposite, -Main, means being governed. However, a clause that expresses direct speech is also categorized as +Main in this study, while being annotated as an object of a communicative verb.

+ccVerb, the property of being a conjunct of a main verbal item, possibly through a chain of coordinations.

+Attributive, the property of being an attribute of a noun. This implies being +Participial.

The clause ranks are defined as conjunctions of these features. The definitions of the clause ranks are summarized in Table 3.

From Ingo's description it is not clear how conjoined clauses should be treated. I have chosen to define complete clauses that are coordinated with a main clause as major, whereas verb phrases that lack an explicit subject are given a category of their own even though they may be coordinated with a main (or subordinate) clause.

There are a few problems in the above definitions with respect to how well they capture the intended concepts. One concerns the distinction between participles on the one hand, and adjectives and nouns on the other. This distinction can be drawn in different ways, but the category participle tends to be a bit overused in LinES, as the basic criterion is a formal one. While participles are more common in English, this tendency is the same for both English and Swedish. Another decision that has effects on the numbers is that verbs with similar meanings may be classified as an auxiliary for one language, but as a non-auxiliary for the other. Also, annotation errors can still be found that affect the classification.

5 Results

Using the clause rank definitions we can simply count the number of clauses at each rank. For the reasons given in the previous section, the figures reported, see Table 4, are not exact but are stable enough to reflect tendencies. They give support to our hypotheses with one exception. In agreement with the hypotheses, Swedish has more finite clauses and, in particular, more minor clauses than English. The number of participial clauses on the English side is more than four times as many as on the Swedish side. But, contrary to the hypothesis, the Swedish side also has more infinitival clauses than the English side.

Table 4 also shows that the tendencies are quite stable across the sub-corpora. The Europarl corpus is slightly divergent, with fewer major clauses on the Swedish side than the English side and almost equal number of infinitival clauses. This is not surprising given that it partly contains parallel translations and has been shown to differ from the other three subcorpora in other respects as well [2].

Clause rank	**Access**		**Europarl**		**Bellow**		**HarryP**		**Sums**	
	En	**Se**	**En**	**Se**	**En**	**Se**	**En**	**Se**	**En**	**Se**
major clause	473	492	598	572	640	655	676	700	2387	2419
minor clause	298	403	261	327	289	371	305	417	1153	1518
coordinated VPs	33	41	14	26	39	72	91	159	177	298
participial clause	273	37	131	37	150	52	217	56	771	182
infinitival clause	210	243	176	172	149	200	138	189	673	804

Table 4: Frequency of clauses at different ranks distributed on sub-corpora.

To see how the different ranks have been treated in translation, we need to exploit the alignment. As the alignment is based on words, it may happen that single words are aligned to word sequences on the other side. Nevertheless, we

Source side clause ranks	**Target side clause ranks**						
	Major	Minor	ccVP	Pcpcl	Infcl	NP	Other
participial clause	78	184	81	56	175	50	149
infinitival clause	83	110	14	1	404	16	45

Table 5: Frequencies of mappings for English clauses of medium rank.

Clause rank with function	Major or minor	ccVP	Pcpcl	Infcl	Other
Adverbial participial clause	51	35	17	10	41
Nominal participial clause	52	5	0	111	48
Modifying participial clause	103	5	17	6	60
Adverbial infinitival clause	38	4	0	66	9
Nominal infinitival clause	90	3	0	137	25
Modifying infinitival clause	39	0	1	79	13

Table 6: Frequencies of mappings depending on grammatical function.

can look at clause governors and their individual images and see to what extent that image includes a clause governor on the other side, and, if so, what rank that clause belongs to. It may also happen that a clause is nominalized in translation, and we include those cases as well. However, the fact that a verb or participle is aligned with a noun does not guarantee that the image is a nominalization; it may, for example, be the result of a single verb being mapped to a complex verb construction such as English *decide* being translated by Swedish *fatta beslutet* ('make the decision').

We focus on participial and infinitival clauses as these are the ones showing the greatest differences in numbers. Data for these two ranks are shown in Table 5. We can see that English participial clauses, when translated into Swedish, yield both clauses of higher rank, and clauses of lower rank. The most common translations are minor clauses and infinitival clauses, while only about 6% of the translations are Swedish participial clauses. The category 'Other' also has many instances, the majority being distributed on prepositional and adjectival phrases, and deletions. Infinitival clauses are sometimes translated into higher ranks, but a large majority, 60%, are translated as infinitival clauses.

The function of the clause has an impact on restructuring. If we divide the participial and infinitival clauses into different groups depending on whether they have an adverbial, modifying or nominal (subject, object, or oblique object) function, we can see that for participial clauses, adverbial and modifying clauses tend to be rendered as complete clauses to a much larger extent than those with nominal functions, as shown in Table 6. For infinitival clauses the function has less impact on the proportion of cases that are rendered as complete clauses.

It can be seen that about 40% of the participial clauses are translated by tensed

clauses or phrases, and that almost 30% also have an overt subject. For English infinitival clauses the corresponding proportions are just below 30%. Even if the numbers for participial clauses on the English side may be exaggerated there are a significant number of instances where human translators select a higher clause rank than the one appearing in the English source and thus, supplying a tense, and a subject or a place-holding subordinator that are congruent with the context. Some examples are given in Table 7.

Mapping	**Clause pair**
pcpcl → minor (Access)	EN: In MS Access 2000 *using ADOX* SE: *När du använder ADOX* i MS Access 2000
pcpcl → major (Bellow)	EN: *Anticipating a difficulty*, I ask the stewardess... SE: *Jag förutser ett problem* och ber flygvärdinnan...
pcpcl → major (Europarl)	EN: ... is also imprudent in *introducing issues not included...* SE: ... *tar* också på ett oförsiktigt sätt *upp frågor som inte förekom...*
pcpcl → minor (Harry P)	EN: ... felt right into the corners *before sweeping the whole lot...* SE: ... kände efter långt inne i hörnen *innan hon sopade ner allt...*
pcpcl → major (Access)	EN: A different layout lets you *calculate and compare...* SE: Med en annan layout *kan du beräkna och jämföra...*
infcl → minor (Harry P)	EN: ..punish himself most grievously *for coming to see you* SE: ..bestraffa sig själv ytterst hårt *för att han hälsat på*

Table 7: Examples of high rank Swedish translations of medium rank English clauses.

6 Conclusions

A dependency-based annotation scheme which notionally distinguishes relations at the clause level and relations at the phrase level can also be used to identify clauses of different ranks. This allows hypotheses as regards restructuring at the clause level in translation to be tested and instances of such changes to be investigated in more detail, whether by human or machine translators.

The LinES data largely confirms the hypothesis that clauses without tense and subjects are more common in English than in Swedish translations. However, in LinES, this is entirely due to participial clauses, while the infinitival clauses are more common on the Swedish side than on the English side. Still, in a sizeable number of cases human translators selects a clause type of higher rank, with material that needs to be congruent with the context. This would seem to pose a problem for current approaches to statistical MT.

References

[1] Ahrenberg, Lars (2007) LinES: An English-Swedish Parallel Treebank. In *Proceedings of The 16th Nordic Conference of Computational Linguistics*, Tartu, Estonia.

[2] Ahrenberg, Lars (2010) Alignment-based profiling of Europarl data in an English-Swedish parallel corpus. Proceedings of the Sixth Conference on Language Resources and Evaluation (LREC'2010), Malta, 19-21 May, 2010.

[3] Baker, Mona (1993) Corpus Linguistics and Translation Studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli (eds.) *Text and Technology*, Amsterdam and Philadelphia: John Benjamins, pp. 233-250.

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

[5] Collins, Michael, Koehn, Philipp, and Kucerova, Ivana (2004) Clause restructuring for statistical machine translation. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, pp. 531 - 540.

[6] Elming, Jakob (2008) Syntactic reordering integrated with phrase-based SMT. In *Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation (ACL-08 SSST-2)*. Columbus, Ohio, USA.

[7] Ingo, Rune (2007) *Konsten att översätta: översättningens praktik och didaktik.* Lund, Studentlitteratur.

[8] Merkel, Magnus and Michael Petterstedt and Lars Ahrenberg (2003) Interactive Word Alignment for Corpus Linguistics. Proceedings of Corpus Linguistics 2003. UCREL Technical Paper No 16.

[9] Saldanha, Gabriela (2009) Principles of corpus linguistics and their application to translation studies research. *Revista Tradumàtica - Traducció i Tecnologies de la Informació i la Comunicació* No. 07.

[10] Tapanainen, Pasi and Timo Järvinen (1997) A non-projective dependency parser. In Proceedings of the 5th Conference on Applied Natural Language Processing, Washington, D.C, April 2007, pp. 64-71, Association for Computational Linguistics.

[11] Vinay, J-P and J. Darbelnet (1977) *Stylistique comparée du français et de l'anglais.* Paris, Didier.