Through a glass darkly

Part-of-speech distribution in original and translated text

Lars Borin and Klas Prütz

Department of Linguistics, Uppsala University

Abstract

In the past, so-called *translationese* has been investigated mainly as a lexical phenomenon, despite suggestions that it also must have a syntactic dimension. In this article, we explore the use of part-of-speech- (POS) -tagged parallel and comparable corpora as one means of investigating translation effects in the syntactic domain. We suggest a method for isolating putative translation effects in the form of over- and underused POS n-grams, which relies upon the existence of tagged comparable corpora for all the investigated languages, and not only for the target language, as in previous investigations. We discuss some of the patterns of overuse which we have found using this method, and some ways in which the method could be used for other investigations.

1 Introduction

Translationese has been characterized as "deviance in translated texts induced by the source language" (Johansson and Hofland 1994, 26). The kind of deviance referred to here is not to be equated with errors in the normal sense, however. Rather, it should reveal itself in 'odd' choices of lexical items and syntactic constructions, which conceivably could be the result of both assimilation and dissimilation with respect to the source language or the source text. Intuitively, the idea that the source language could influence the target language in this way seems plausible to anybody who has struggled to convey his message in a recalcitrant foreign language.[1]

Intuition is a good servant but a bad master, and should be backed up by principled empirical investigation. Translationese could be argued to fall under the more general heading of language contact phenomena, which have been the subject of linguistic research for a long time (e.g. Weinreich 1953; Thomason and Kaufman 1988; Saxena 1997). On the other hand, language contact researchers investigate above all fairly obvious and eminently noticeable changes in the linguistic system or a subsystem of a language. Translationese, on the other hand, should, by its nature, manifest itself by more subtle means than the methodology of language contact research normally is equipped to handle. It ought to be seen, above all, in deviant patterns of usage, i.e. it should be eminently suited for investigation by the methods of corpus linguistics.

No surprise, then, that it is among corpus linguists that we find many students of the linguistic correlates of translationese. However, their studies mostly confine themselves to lexical phenomena (e.g. Gellerstam 1985; Ebeling 1998; Johansson

[1] This is only an analogy, of course, since translators normally translate into their native language.

forthcoming), but translation effects should be noticeable in other linguistic domains as well, e.g. in syntax. Gellerstam (1985, 94) in his mainly lexical study states that "no doubt there are also syntactic fingerprints in translations", but does not elaborate on the matter beyond giving a single example from his corpus.

The reason for the predominance of lexical studies of translationese is in all probability that the tools are readily available for carrying out lexical investigations on languages with none or insignificant morphology (to draw a somewhat arbitrary line: those where most inflected parts of speech have less than 10 forms in their inflectional paradigms). These tools are concordancers, possibly with some statistical processing capabilities, and sentence aligners. Using them, you simply create monolingual and bilingual concordances for the lexical items that you are interested in, and then analyze this data in time-honored linguistic fashion. In a somewhat more exploratory manner, you can produce frequency lists of word types, word bigrams, trigrams, etc., and compare, e.g., vocabulary size and spread between translated and original texts of the same general type, as well as collocational patterns. This is done on the level of text words, however, i.e. inflected forms, and not that of lemmas or lexemes, which is what you normally would be interested in in a study of this type.

In order to investigate syntactic phenomena, we need texts which have been provided with some kind of syntactic markup, or tagging. At the very least, they should be tagged for part-of-speech (POS), which actually is something of a misnomer, because POS tags are often actually fairly detailed morphological descriptions. POS-tagged corpora are still hard to find, except for a few languages and a few text types, and taggers that you can use yourself are still much less common than concordancers, and for obvious reasons much more language-dependent. Even more useful for syntactic studies of translation phenomena would be parsed text material, of course, but this is harder still to come by.

In stylometric studies, such as authorship attribution, tagged corpora have been used for quite some time (e.g. Kjetsaa et al. 1984). In the study of second and foreign language learning, more specifically the study of interlanguage (see Selinker 1992), there have recently been studies in which learner corpora (see Granger 1998) have served as basis for investigations of POS n-gram differences between native and learner English (Aarts and Granger 1998; Berglund and Prütz 1999), with a method which is very similar to the one proposed here, but with the very important difference that tagged L1[2] texts are not used.

The present work represents an attempt to move studies of translation effects into the syntactic arena. It benefits from the corpus collection and tagging work done as part of the ETAP project,[3] and will benefit also from the work with sen-

[2] I.e., the language learners' native language, Dutch, Finnish and French in Aarts and Granger's (1998) investigation.

[3] ETAP is the acronym of the project title "Etablering och annotering av parallellkorpus för igenkänning av översättningsekvivalenter" (in English: "Creating and annotating a parallel corpus for the recognition of translation equivalents"). This project is a part of a joint research programme between the universities in Stockholm and Uppsala, "Translation and Interpreting – A Meeting between Languages and Cultures" financed by the Bank of Sweden Tercentenary Foundation (Riksbankens Jubileumsfond); see <http://www.translation.su.se>.

tence and word alignment done in the same project.

We will investigate whether differences in POS n-gram occurrences are indicative of translation effects. For our investigation, we have used two ETAP corpora, representing news text translated into English from Swedish, together with the parallel Swedish original, and two publicly available corpora of original English, the Flob and Frown corpora. The corpora are described in more detail in the next section. In section 3, we describe how the tagging of the corpora was done, as well as the tagsets used. Section 4 is devoted to finding the differences between original and translated English text with regard to POS n-grams, using a method where both L1 and L2 POS-tagged corpora are used. Section 5 contains a discussion of our findings, and in section 6, we sum up and look ahead.

2 The corpora

For this investigation, four corpora were used, all representing the text type news text. The corpora are

(1-2) The Flob corpus is a corpus of British English compiled in the 1990's at Freiburg University with the same composition as the well-known Lancaster–Oslo–Bergen (LOB) corpus of British English. We used the parts of the corpus marked as "press, reportage", being the category which we deemed most similar in content and style to the IVT1 corpus (see below).[4]

Similarly, the Frown corpus is a more recent version of the Brown corpus of American English, also compiled at Freiburg University. Here, too, we used the press/reportage parts of the corpus.

(3-4) The Swedish and English portions of the parallel newspaper text corpus IVT1 of the ETAP project. *Invandrartidningen* (IVT) is a periodical for immigrants in Sweden, appearing in 40 issues annually, in 8 language versions, Arabic, Bosnian–Serbian–Croatian, English, Finnish, Persian, Polish, Spanish, and simplified Swedish. The IVT1 corpus is made up of about half a year's worth of issues of IVT in five languages: the Swedish original (which is not published as such, but only used for translation into the other languages, including 'translation' into simplified Swedish),[5] Bosnian–Serbian–Croatian, English, Polish, and Spanish.

In Table 1, we give some statistics for the four corpora.

[4]The Flob and Frown corpora also have two other "press" categories, viz. "editorial" and "review", but as *Invandrartidningen* is a fairly pure news publication, we deemed it better to leave these out. Furthermore, the "reportage" category in Flob and Frown turned out to contain almost exactly the same amount of text as IVT1 (see Table 1).

[5]We are grateful to the *Invandrartidningen Foundation* and the editor-in-chief of *Invandrartidningen*, Dag Zotterman, who graciously made electronic and paper copies of the periodical available to us, as well as the Swedish original manuscript material from which all translations were made.

Table 1: Word and POS statistics for the four corpora

	Flob	Frown	IVT/EN	IVT/SE
tokens	98855	101319	119779	97339
word 1-grams (= types)	14625	14741	12702	15890
word 2-grams	62828	63941	62988	60343
word 3-grams	90578	92276	101843	87763
POS tag 1-grams	30	30	30	36
POS tag 2-grams	584	618	608	835
POS tag 3-grams	4679	5163	5241	6359
POS tag 4-grams	18471	20281	20939	22756
POS tag 5-grams	42871	46474	48930	48540

3 Tagging the corpora

All four corpora were tagged with a Brill tagger (Brill 1992). The Swedish tagger was trained on another ETAP subcorpus, the SGP corpus of political texts, on newspaper texts (from the local daily *Upsala Nya Tidning*, graciously made available for our use by the SCARRIE project), and on fiction texts from the Stockholm Umeå Corpus (Ejerhed and Källgren 1997), using a tagset devised to be compatible with the morphological descriptions in SVE.UCP (Prütz forthcoming; Sågvall Hein 1988; Sågvall Hein and Sjögreen 1991). The English tagger was trained on the written part of the BNC Sampler (Burnard 1999), using the BNC tagset (Leech and Smith 1998). See the Appendix for a listing of the two tagsets.

The Swedish Brill tagger has been tested on a held-out subset of the SGP corpus and the accuracy is estimated to 95.7 per cent correct tags. The tagger trained on the BNC Sampler text was tested using text from the Uppsala Student English corpus (USE; Axelsson 2000; Axelsson and Berglund forthcoming), giving an estimated accuracy of 96.7 per cent correct tags.

For the purposes of this investigation, both tagsets were reduced after the texts were tagged, the English set from 145 to 30 tags and the Swedish one from 151 to 36 tags (the reduced tagsets are listed and compared in the appendix). This was done for two reasons.

First, earlier work has indicated that training and tagging with a large tagset, and then reducing it, not only improves tagging performance, but also gives better results than training and tagging only with the reduced set. Prütz's (forthcoming) experiment with the Swedish Brill tagger and the same full and reduced tagsets as those used here gave an increased accuracy across the board of about two percentage points from tagging with the large tagset and then reducing it, compared to tagging with the full set. Tagging directly with the reduced set resulted in a lower accuracy, by a half to one percentage point, depending on the lexicon used.

Second, coarse-grained tagsets are more easily comparable than fine-grained ones even for such closely related languages as Swedish and English (Borin 2000, forthcoming).

4 POS n-gram differences between original and translated English text

The main hypothesis which inspired the work reported here is that 'translationese' is not confined to the lexical level, which has been the one normally investigated in works on translationese (see above in section 1). Further, we believe that distributional differences in POS n-grams (with n ranging from 1 to 5 in our investigation) may turn out to be indicative of translation effects in the syntactic domain.

First, we will look at the simplest case, that of POS unigram (n=1) frequencies. In Table 2, POS unigram frequencies for the four corpora are shown, excluding tags for punctuation. On the surface of it, POS unigrams do not seem very promising for illustrating translation effects. On the contrary, the three English corpora are very similar in their POS distribution, and different from the Swedish text in roughly the same ways. The differences include a significantly[6] greater number of NN (common noun) tags for all English texts compared to the Swedish corpus. We really do expect this to be the case, due to the way the orthographies of the languages work; in English, (noun–noun) compounds are normally written as two (or more, if one of the parts is in itself a compound) orthographic words, while in Swedish—just as in German—the parts are written together as one orthographic word (examples from IVT/SE – IVT/EN):

vapenexport	—	nattåg	—	polisrazzia
arms exports	—	night trains	—	police raid

For the same reason, we expect—and find as well—significantly more T (determiners, including articles) tags in English than in Swedish, because of the definite article being written separately in English but as part of the noun (an inflectional suffix) in Swedish.

But there are also more intriguing differences, less easily explained by differences in language structure or orthographic conventions. One such difference concerns the POS tag R (adverb), where Swedish has significantly more R tag instances than any of the three English corpora, among which IVT/EN has the highest number of R tags (although not significantly more than Flob or Frown). Even though we have not looked at the details of this case (are there many more adverb lemmas, or simply more of the same ones?), we are reminded of the following observation on the differences between English and Kalam, a language of the New Guinea Highlands:

> The special features of Kalam event-reports first surfaced as a language-learning difficulty. I had been living in the Upper Kaironk for a couple of months and had learnt to converse, hesitantly, about a range of familiar subjects. I noticed that bystanders, who were fond of repeating to others nearby what I said (even if the others could hear

[6] Significance testing was done using the Mann-Whitney test, following Kilgarriff's (to appear) suggestion that this is a more suitable test for determining which units are used most differently in two text corpora than, e.g., the χ^2 test. The significance level used throughout was $p \leq 0.025$. See further section 4 below.

Table 2: POS unigram frequencies in the four corpora (excluding punctuation)

Flob		Frown		IVT/EN		IVT/SE		rank
NN	21933	NN	21919	NN	22020	NN	20196	1
I	10467	I	9703	I	9726	V	11799	2
V	8536	V	8462	V	8722	I	10675	3
T	8126	T	7737	T	7601	P	7219	4
NC	6553	A	6898	A	6576	R	6750	5
A	6444	NC	6604	P	6567	A	5911	6
P	5646	P	5451	R	4839	C	4672	7
R	4582	R	4308	C	4238	NC	4063	8
C	4116	C	4262	NC	3895	VI	3877	9
VI	2888	VI	2808	VI	3443	T	3355	10
K2	2706	M	2369	K2	2525	F	1369	11
M	2263	K2	2088	M	2302	E	1319	12
E	1676	E	1586	E	1680	VS	1239	13
K1	1629	K1	1550	K1	1429	Q	1172	14
P$	1257	P$	1118	P$	1241	M	1150	15
$	184	$	593	$	565	P$	844	16
O	53	O	91	O	82	NN$	686	17
S	29	S	52	S	77	K2	565	18
X	25	X	21	X	27	NC$	224	19
						G	187	20
						L	162	21
						K1	124	22
						O	54	23
						S	48	24
						X	24	25
						VK	18	26

> perfectly well), often added details to my utterances. For instance, if someone asked, “Where’s Kiyas?” (the young man who was my chief informant) and I answered, “He’s in his garden”, a bystander might say, “He said ‘Kiyas has gone to Matpay to work in his garden. He’ll be back later’, he said”.
>
> After a while it dawned on me that these elaborations were not just imaginative creations of individuals but followed a consistent pattern. People were editing my utterances, supplying information that I should have given in the first place to make my utterance complete. (Pawley 1993, 109)

It could well be that Swedish newswriters feel a greater need to supply where, when and how events took place than their English counterparts would.[7] It is a

[7]It could also be that in English, the preference is for adverbials in the form of e.g. prepositional phrases, rather than simple adverbs. We would need at least a syntactically parsed corpus in order to

different question whether this tendency in news text reflects a deeper difference in the genius of the two languages, as Pawley claims for English and Kalam in the passage just quoted, or whether it points to a difference in preferred news text style in the two languages, Swedish preferring a more colloquial (or concrete) style and English a more formal (or abstract) language.

Similarly to the adverbs, there is a significantly higher incidence of infinitives (VI) and pronouns (P) in the IVT/EN text, compared to Flob and Frown, corresponding to even higher figures for the IVT/SE text. These differences could reflect translation effects, as follows.

In the case of the infinitives, there is a readily available structural factor which could account for the effect: Linguistic system constraints force you to use the infinitive in most dependent non-finite clauses in Swedish, whereas in English there is also the present participle/gerund (K1 in the reduced tagset used here) available as an alternative, depending on the main clause. The translation effect in this case would be seen as a tendency in the translator to translate (obligatory) infinitives with (optional) infinitives, choosing an appropriate main clause form for this to be possible.

As for the pronouns, however, there is no such structural explanation that leaps to mind. If the Swedish news style is more colloquial than its English counterpart, as conjectured above, the higher incidence of pronouns could be a mark of this. In spoken English, pronouns are much more frequent than in the written variety. Thus, in the London-Lund Corpus of spoken English, pronouns are actually more frequent than nouns (Altenberg 1990, 185). In academic writing by Swedish university students, i.e. advanced learners of English, there is also an overuse of pronouns (Axelsson and Berglund forthcoming; Berglund and Prütz 1999). This fact should be seen in the light of observations about how well Swedish university students of English master the colloquial registers of the language, but have less training in the more formal registers, and consequently display an excessively colloquial style in their formal written production (Ohlander 1995).

We now turn to an investigation of n-grams where n>1. There are many more 2-, 3-, etc. grams than 1-grams (see Table 1), and it is not feasible to do this investigation manually. Instead, we followed the procedure described below, which is logically divided into a (computationally less demanding) *hypothesis generation* stage and a *hypothesis testing* stage. The intention is to identify n-grams evidencing putative translation effects in the hypothesis generation stage, and then subject these to significance testing. Thus, hypothesis generation was done as follows.

1. First all texts were tagged as described in section 3 above.

2. POS n-grams were extracted from the texts, and sorted in order of decreasing frequency. All frequencies were normalized; figures shown in Table 3 are frequency/100000 tokens. The rank of the n-grams was defined to be in-

investigate this hypothesis (see section 6). Against this conjecture we may adduce the fact that the frequency of prepositions is roughly the same in all our corpora; in fact, of the four corpora, IVT/SE has both the most adverbs and the greatest number of prepositions (see Table 2).

versely related to their frequency, so that the item with the highest frequency gets rank 1, etc.[8]

3. The Flob n-gram ranking was then compared to the others. We used Flob as the standard against which the other texts were compared because the IVT/EN texts follow British English most closely in their orthography, vocabulary, etc. Thus, for each of the text pairs Flob–Frown, Flob–IVT/EN, and Flob–IVT/SE, we produced a thresholded rank difference list, using a (heuristically chosen) threshold of 30, i.e. the rank difference must be 30 or greater for it to count as a difference. In the rank comparisons, a positive number means that the n-gram in question has a lower frequency in the other text, and a negative number that it has a higher frequency.

4. The difference lists were then processed as follows. First, we ran the comparisons of Flob with IVT/EN and IVT/SE through a small program which kept only rank differences which IVT/EN and IVT/SE had in common in comparison with Flob, i.e. differences with the same sign, hence both denoting either higher or lower rank.

5. After this, we did the exact opposite, but with Frown as the comparison, i.e. we discarded from the result all rank differences common to the comparisons of Flob with Frown and with IVT/EN. Thus, Frown was used as a control, as it were, helping us avoid ascribing rank differences to translation effects, when they are in fact simply an effect of the normal variation found in the investigated text type. In this way, 2 2-grams (of 29), 36 3-grams (of 98), 14 4-grams (of 72), and 1 5-gram (of 9) were eliminated.

6. Finally, certain n-grams were removed from the resulting lists. All n-grams containing the tag NC (proper noun) were discarded, since we believe that a higher or lower relative propensity of proper nouns is not a distinguishing trait in translationese. We also discarded all n-grams containing punctuation, except those having a full-stop as their first or last tag (but no punctuation tags elsewhere in the sequence). The motivation for this is is less well-founded, but let us simply say that we wish to limit ourselves, at least for the time being, to looking at clause-internal syntax darkly mirrored in a POS tagging of a text.[9] The elimination of punctuation tags resulted in a further

[8]Ties get the same rank, and there are no unfilled positions in the ranking. Thus, the frequency distribution 8, 7, 6, 6, 6, 2, 2, 1, 1, 1, 1, 1 would result in 5 ranks, numbered 1–5, and having 1, 1, 3, 2, and 5 members, respectively. The highest-numbered ranks in the corpora—i.e., for the n- grams with frequency 1—were as follows.

	Flob	Frown	IVT/EN	IVT/SE
2-grams	211	219	221	223
3-grams	242	255	254	229
4-grams	157	156	169	149
5-grams	99	94	97	83

[9]Of course, at the same time we eliminate e.g. commas functioning as coordination conjunctions,

reduction of the number of n-grams. A total of 20 2-grams, 39 3-grams, 36 4-grams, and 6 5-grams were eliminated in this step.

In the following stage of the investigation, the hypothesis testing was done using the Mann-Whitney (or U) test for each of the surviving n-grams (see Kilgarriff to appear), and we kept only those n-grams simultaneously showing a significant difference between the two corpus pairs Flob–IVT/EN and Flob–IVT/SE above the 97.5% level ($p \leq 0.025$) for a directional test (since we did know the expected direction of the difference). In Table 3, we see the n-grams remaining after removal of non-significant (in the sense just described) differences (6 3-grams and 5 4-grams were eliminated by the test).

5 Discussion

In this preliminary study, we will limit ourselves to discussing a small number of representative cases. Translation effects should in principle manifest themselves as both overuse and underuse of syntactic constructions, just as the case is in foreign language learners' interlanguage (Aarts and Granger 1998). Here, we will only discuss cases of overuse in our material—i.e. those where the rank difference ($\Delta rank$) is a negative number in Table 3—but we hope to be able to return to the equally interesting cases of underuse at a later time.

There are some n-grams indicating that there are more verb-initial sentences in IVT/EN than in the two original English corpora (the 2-gram ". V" and the 3-gram ". V P"). This is probably not due primarily to a translation effect, however. Rather, it reveals a difference in text type composition among the corpora. The IVT corpora contain a fair amount of text from the section "Letters from the readers" in the periodical. The language of this section differs from that of the rest of the corpus, e.g. in containing a large amount of direct questions. Hence the many verb-initial sentences, characteristic of Swedish yes-no questions:

Examples from the "Readers' letters" section of IVT, issue 20, 1997 (sentence-initial V underlined)

IVT/SE:	IVT/EN:
Måste djuren sitta i karantän? Jag har en liten hund kvar i USA, kan jag ta hit den? Måste den sitta i karantän?	Must our animals be kept in quarantine? I have a small dog in the United States. Can I bring it here? Must it be kept in quarantine?

On the other hand, an example of what seems to be a real instance of translationese syntax is the overuse of preposition-initial sentences in IVT/EN, as seen in the 2-gram ". I" (full-stop–preposition). There is no difference in preposition (1-gram) frequency among the four texts (see Table 2), however. Thus, we seem to be

i.e. clause-internally. We also do not wish to claim that rules of orthography, such as the use of punctuation, cannot be subject to translation effects. We are simply more interested in syntax more narrowly construed. The reason for keeping leading and trailing full-stops is that a full-stop is an unambiguous sentence (and clause) boundary marker, thus permitting us to look at POS distribution at sentence (and some clause) boundaries.

Table 3: Remaining significantly different n-grams ($p \leq 0.025$) after filtering through Frown and removal of sequences containing NC and punctuation tags

Flob freq	Δrank	2-gram	Flob freq	Δrank	5-gram
366	71	T M	60	36	NN I T M NN
267	-38	. I	25	-30	NN I P$ NN .
176	-58	. R			
64	47	P$ M			
62	-65	P VI			
60	-52	C VI			
21	-95	. V			
Flob freq	Δrank	3-gram	Flob freq	Δrank	4-gram
209	93	T M NN	116	78	I T M NN
204	105	I T M	103	60	NN E VI T
189	49	NN A NN	102	44	A NN E VI
162	-42	I P NN	101	74	NN I T M
161	92	V P V	101	-36	I A NN .
142	-38	P NN V	94	65	A NN NN NN
114	53	A NN E	88	57	NN NN NN I
109	-44	. T A	84	62	K1 T NN I
106	-43	P$ NN .	82	-51	. P V R
95	50	M NN C	78	43	I T A A
87	-50	P NN .	64	43	P V P V
57	-54	. P NN	63	-37	I P$ NN .
51	39	T M A	60	49	V T NN V
31	-60	NN C VI	50	37	K2 I NN NN
20	-53	. I P	39	-50	. P V VI
17	-121	V P VI	39	-38	NN V R A
2	-69	. V P	29	-32	. P NN V

confronted with a difference between Flob and IVT/EN which mirrors a difference between Flob and IVT/SE, i.e. a putative translation effect.

We know that Swedish is more liberal than (written standard) English when it comes to allowing constitutents other than the subject in the sentence-initial position of declarative sentences. Frequently, you will find quite heavy adverbials—which are often prepositional phrases—or prepositional objects in this position, while in English, although certainly possible, this construction is less preferred than in Swedish:

Examples from IVT, issue 19, 1997 (sentence-initial prepositions underlined)

IVT/SE:	IVT/EN:
För att bli svensk folkmusiker tog Ale en ovanlig omväg. I Malmö, där han bodde, blev han bekant med en invandrad grekisk musiker som satte en bouzouki i händerna på Ale.	In becoming a Swedish folk musician, Ale Möller took a strange path. He lived in Malmö and got to know an immigrant Greek musician who placed a bouzouki in his hands.

Possibly the same feature of Swedish syntax lies behind the overuse of sentence-initial adverbs (the 2-gram ". R") in IVT/EN, as compared to Flob. We have already seen that adverbs by themselves are overused in IVT/EN, and now it seems that quite a few of those extra adverbs end up in sentence-initial position.

The overuse of pronouns noted earlier is further reflected in the 4-grams ". P V R" (full-stop–pronoun–finite verb–adverb) and ". P V VI" (full-stop–pronoun–finite verb–infinitive), where—as revealed by an inspection of the actual text word sequences—the P corresponds in all instances to pronominal subjects (in the form of personal, demonstrative, or expletive pronouns). In the same way, the P in the (clause-internal) 3-gram "V P VI" (finite verb–pronoun–infinitive) corresponds in practically all cases to a pronominal subject or object in the text.

Finally, the higher use of the 2-gram "C VI" (conjunction–infinitive) and the 3-gram "NN C VI" (common noun–conjunction–infinitive) could be seen as further confirmation of the conjecture made in section 4 above, that the translator of IVT tends to carry over Swedish infinitive clauses, even when an English verb form in *-ing* perhaps should have been the preferred choice.

6 Conclusions and outlook

Our results, although of a preliminary nature, are encouraging. It seems that we are able to tease out some interesting syntactical traits of so-called translationese, using POS tagged corpora and the method described in section 4 above.

To be more precise, we have shown that there are significant differences in the distribution of POS n-grams that IVT/EN and IVT/SE share when contrasted against Flob, that could be indicative of a translation effect in IVT/EN.

Here, we have to put in a caveat: We cannot be absolutely certain that we have, in fact, produced firm evidence for translationese in IVT/EN, as the differences that we have seen could, in principle, be due to non-linguistic differences between the corpora, e.g. differences in content or topic. We could be dealing with different sublanguages, as it were, with slightly different syntactic profiles. One such difference has turned up already, namely the "Letters from the readers" material present in IVT, but not in Flob (discussed in section 5 above). We believe that the IVT and Flob corpora are otherwise comparable as to content and topic, but only further more detailed examination can show whether this belief is justified or not.

The method naturally lends itself to a working mode where we go from linguistic abstractions, i.e. POS n-grams, to increasingly concrete cases, i.e. via more specific—or longer—n-grams, to sequences of text words corresponding to particular POS n-grams. It is important to note that hypotheses are formulated on

the basis of the abstractions, but checked out with the help of increasingly less abstract representations. Thus, the ability to create—manually or partly or fully automatically—linguistic abstractions of texts is a necessary prerequisite for any linguistic investigation. Corpus linguistics brings to this process the possibility of making automatic such abstractions of large amounts of text, thus enabling us to discern subtle patterns of usage in a more 'objective' way than previously (cf. Grefenstette forthcoming).

If we compare our method to that used by Aarts and Granger (1998) in their investigation of interlanguage in language learning, we may note that we use an L1 POS tagged corpus (IVT/SE), in order to correlate the differences between L2 and L1 English to those between L1 English and L1 Swedish, while Aarts and Granger use only L2 learner English texts, comparing them with L1 English texts, but not with texts in the learners' native languages (Dutch, Finnish and French). No doubt, there were practical reasons for working only with the target language, but there is also a school of thought in second and foreign language learning research which holds that interlanguage goes through more or less the same stages, regardless of the learner's native language (see Lightbown and Spada 1993), so that there would, strictly speaking, be no need to look at anything but the target language. We still believe, however, that contrastive factors are important both in the case of translators and second language learners.

An obvious further development of the investigation reported here would be to look at parts-of-speech in certain positions, i.e. to look at more abstract 'meta-patterns', defined e.g. by regular expressions over POS sequences. Thus, it would be interesting, for instance, to investigate 3-grams with sentence punctuation[10] in the first position and finite verbs in third position, to see whether the obligatory V2 structure of Swedish influences English translations from Swedish. We have seen that one aspect of V2 structure—viz. that there is a greater tendency for other constituents than the subject to occupy the first sentence position—seems to be a feature of Swedish–English translationese, but by abstracting even more away from the text in the way sketched here, we could possibly see whether this influence goes further.

Of course, we know that the "2" in V2 refers to a constituent position, and not to a word position, but with a POS-tagged corpus, word positions are all that we have. The hypotheses which must be assumed in order for the procedure just outlined to yield valid results are (1) that phrases consisting of single words—e.g. NP's consisting of single nouns—are frequent, and (2), that their relative frequency is approximately the same in both original and translated text.

This is not, on the whole, a good assumption—at least not the second part of it—which brings us to another natural continuation of the work reported here. Even though POS-tagged texts allow us to make interesting observations about the differences between original and translated language, parsed text would be even more useful, but such texts and publicly available parsers for a range of languages are much harder to come by than POS taggers.

[10] Sentence punctuation being represented also by other tags in addition to the full-stop tag (see the Appendix) seen in the n-grams discussed in sections 4 and 5.

Finally, we would like to try out the method described here (or a refinement of it) on learner language as well, comparing the results we would get with those achieved by Aarts and Granger (1998) and Berglund and Prütz (1999). This would be a first step toward creating error tagged learner corpora (cf. Dagneaux et al. 1998) for at least English and Swedish, with the ultimate aim of using these resources in intelligent computer-assisted language learning (ICALL) applications.

Acknowledgements

We wish to thank the participants of the joint Stockholm–Uppsala Translation Programme Summer Seminar 2000 and two anonymous reviewers for their insightful and constructive comments on successive versions of this text.

References

Aarts, J. and Granger, S.(1998), Tag sequences in learner corpora: a key to interlanguage grammar and discourse, *in* S. Granger (ed.), *Learner English on Computer*, Longman, London, pp. 132–141.

Altenberg, B.(1990), Spoken English and the dictionary, *in* J. Svartvik (ed.), *The London–Lund Corpus of Spoken English. Description and Research*, Lund University Press, Lund, pp. 177–191.

Axelsson, M. W.(2000), USE – the Uppsala Student English corpus: an instrument for needs analysis, *ICAME Journal* **24**, 155–157.

Axelsson, M. W. and Berglund, Y.(forthcoming), The Uppsala Student English corpus (USE): a multi-faceted resource for research and course development, *in* L. Borin (ed.), *Parallel Corpora, Parallel Worlds*.

Berglund, Y. and Prütz, K.(1999), Tagging a learner corpus – a starting point for quantitative comparative analyses, Presentation at the KORFU '99 Symposium, Växjö University, Sweden.

Borin, L.(2000), Something borrowed, something blue: Rule-based combination of POS taggers, *Second International Conference on Language Resources and Evaluation, Proceedings, Vol. 1*, ELRA, Athens, pp. 21–26.

Borin, L.(forthcoming), Alignment and tagging, *in* L. Borin (ed.), *Parallel Corpora, Parallel Worlds*.

Brill, E.(1992), A simple rule-based part-of-speech tagger, *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento.

Burnard, L. (ed.)(1999), *Users Reference Guide for the BNC Sampler*, Published for the British National Corpus Consortium by the Humanities Computing Unit at Oxford University Computing Services, February 1999. Available on the BNC Sampler CD. See <http://info.ox.ac.uk/bnc/>.

Dagneaux, E., Denness, S. and Granger, S.(1998), Computer-aided error analysis, *System* **26**, 163–174.

Ebeling, J.(1998), Contrastive linguistics, translation, and parallel corpora, *META*.

Ejerhed, E. and Källgren, G.(1997), *Stockholm Umeå Corpus version 1.0, SUC 1.0*, Department of Linguistics, Umeå University.

Gellerstam, M.(1985), Translationese in Swedish novels translated from English, *in* L. Wollin and H. Lindquist (eds), *Translation Studies in Scandinavia. Proceeedings from the Scandinavian Symposium on Translation Theory (SSOTT) II, Lund 14–15 June, 1985*, pp. 88–95.

Granger, S. (ed.)(1998), *Learner English on Computer*, Longman, London.

Grefenstette, G.(forthcoming), Multilingual corpus-based extraction and the Very Large Lexicon, *in* L. Borin (ed.), *Parallel Corpora, Parallel Worlds*.

Johansson, S.(forthcoming), Towards a multilingual corpus for contrastive analysis and translation studies, *in* L. Borin (ed.), *Parallel Corpora, Parallel Worlds*.

Johansson, S. and Hofland, K.(1994), Towards an English–Norwegian parallel corpus, *in* U. Fries, G. Tottie and P. Schneider (eds), *Creating and Using English Language Corpora*, Rodopi, Amsterdam, pp. 25–37.

Kilgarriff, A.(to appear), Comparing corpora, *International Journal of Corpus Linguistics*. References here are to ms available on the WWW: <http://www.itri.bton.ac.uk/~Adam.Kilgarriff/ijcl.pdf>.

Kjetsaa, G., Gustavsson, S. and Beckman, B.(1984), *The Authorship of The Quiet Don*, Solum, Oslo.

Leech, G. and Smith, N.(1998), *The Automatic Tagging of the British National Corpus (Information to be used with the BNC Sampler Corpus)*, UCREL, Lancaster University.

Lightbown, P. M. and Spada, N.(1993), *How Languages are Learned*, Oxford University Press, Oxford.

Ohlander, S.(1995), Variation och standard inom universitetsundervisningen i engelsk grammatik, *in* L.-G. Andersson and F. Börjeson (eds), *Språkundervisning på universitetet. Rapport från ASLA:s höstsymposium, Göteborg, 11–13 november 1993*, ASLA, Uppsala, pp. 117–133.

Pawley, A.(1993), A language which defies description by ordinary means, *in* W. A. Foley (ed.), *The Role of Theory in Language Description*, Mouton de Gruyter, Berlin, pp. 87–129.

Prütz, K.(forthcoming), Part-of-speech tagging for Swedish, *in* L. Borin (ed.), *Parallel Corpora, Parallel Worlds*.

Sågvall Hein, A.(1988), Towards a comprehensive Swedish parsing dictionary, *Studies in Computer-Aided Lexicology*, Almqvist & Wiksell International, Stockholm, pp. 268–298.

Sågvall Hein, A. and Sjögreen, C.(1991), Ett svenskt stamlexikon för datamaskinell morfologisk analys, *in* M. Thelander et al (eds.), *Svenskans beskrivning 18*, Lund University Press, Lund.

Saxena, A.(1997), Internal and external factors in language change. aspect in Tibeto-Kinnauri, *Technical Report 32*, Department of Linguistics, Uppsala University.

Selinker, L.(1992), *Rediscovering Interlanguage*, Longman, London.

Thomason, S. G. and Kaufman, T.(1988), *Language Contact, Creolization, and Genetic Linguistics*, University of California Press, Berkeley.

Weinreich, U.(1953), *Languages in Contact*, Mouton, The Hague.

Appendix: Reduced English and Swedish tagsets

SE-R	EN-R	description	examples
–	–	dash	–
!	!	exclamation mark	!
"	"	quotes	”
(	(	left bracket	(
)	)	right bracket	(
,	,	comma	,
.	.	full-stop	.
	...	ellipsis	...
:	:	colon	:
;	;	semicolon	;
?	?	question mark	?
	$	genitive clitic	’
A	A	adjective	röd, red
C	C	conjunction	och, that
E	E	infinitive mark	att, to
F		numeric expression	16
G		abbreviation	d.v.s.
I	I	preposition	på, on
K1	K1	present participle	seende, eating
K2	K2	past participle	sedd, eaten
L		compound part	hög-
M	M	numeral	två, two
NC	NC	proper noun	Eva, Evelyn
NC$		proper noun, genitive	Åsas
NN	NN	noun	häst, goat
NN$		noun, genitive	tjuvs
O	O	interjection	bu, um
P	P	pronoun	vi, we
P$	P$	pronoun, possessive or genitive	vår, our
Q		pronoun, relative	som
R	R	adverb	fort, fast
S	S	symbol or letter	G
T	T	determiner	en, the
V	V	verb, finite	såg, ate
VI	VI	verb, infinitive	se, eat
VK		verb, subjunctive	såge
VS		verb, supine	sett
X	X	unknown or foreign word	