New wine in old skins?
A corpus investigation of L1 syntactic transfer in learner language

Lars Borin[*] and Klas Prütz[**]
[*] Natural Language Processing Section, Department of Swedish, Göteborg University
[**]Centre for Language and Communication Research, Cardiff University

This article reports on the findings of an investigation of the syntax of Swedish university students' written English as it appears in a learner corpus. We compared part-of-speech (POS) tag sequences (being a rough approximation of surface syntactic structure) in three text corpora: (1) the Uppsala Student English corpus (USE); (2) the written part of the British National Corpus Sampler (BNCS); (3) the Stockholm Umeå Corpus of written Swedish (SUC). In distinction to most other studies of learner corpora, where only the target language (L2) as produced by native speakers has been compared to the learners' interlanguage (IL), i.e. their version of L2 (as in the work on the International Corpus of Learner English, ICLE), we add a comparison with the learners' native language (L1) as produced by native speakers. Thus, we investigated differences in the frequencies of POS n-grams between the BNCS (representing native L2) on the one hand, and the USE (representing IL) and SUC (representing native L1) corpora on the other hand, the hypothesis being that significant common differences would reflect L1 interference in the IL, in the form of underuse or overuse of L2 constructions. This makes our study not only one of learner language, or IL in general, but of specific L1 interference in IL. We compare the results of our study to methodologically similar learner corpus research by Aarts and Granger, as well as to our own earlier investigation of English translated from Swedish.

1. Introduction

An important strand of inquiry in second language acquisition (SLA) research is that devoted to the investigation of language learners' successive approximations of the target language, referred to as *interlanguage* (IL) in the SLA literature. Similarly to the practice in other kinds of linguistic investigation, SLA researchers are concerned with empirical description of various kinds of interlanguage, with discovering correlations between traits in interlanguage and features of the language learning situation, with explaining those correlations, and finally with the practical application of the knowledge thus acquired to language pedagogy.

The features of language learning situations which have at one time or another been claimed to influence the shape and development of IL are the following (based on Ellis 1985: 16f):

1. Situational factors (explicit instruction or not; foreign vs. second language, etc.)
2. Linguistic input
3. Learner differences, including learner's L1
4. Learner processes

In this paper, we will be concerned mainly with factor (3), and more specifically with the influence of the learner's L1 on her IL. The phenomenon that features of the learner's native language are "borrowed" into her version of the target language – the IL – is referred to as *transfer* in the SLA literature. Transfer could in principle speed up language learning, if L1 and L2 are similar in many respects, but the kind of transfer which understandably has been most investigated is that where the learner transfers traits which are not part of the L2 system (*negative transfer* or *interference*).

Interference and other features of IL have long been studied by so-called *error analysis* (EA), where language learners' erroneous linguistic output is collected. Traditional EA suffers from a number of limitations:

- *Limitation 1*: EA is based on heterogeneous learner data;
- *Limitation 2*: EA categories are fuzzy;
- *Limitation 3*: EA cannot cater for phenomena such as avoidance;
- *Limitation 4*: EA is restricted to what the learner cannot do;
- *Limitation 5*: EA gives a static picture of L2 learning.

(Dagneaux et al. 1998: 164)

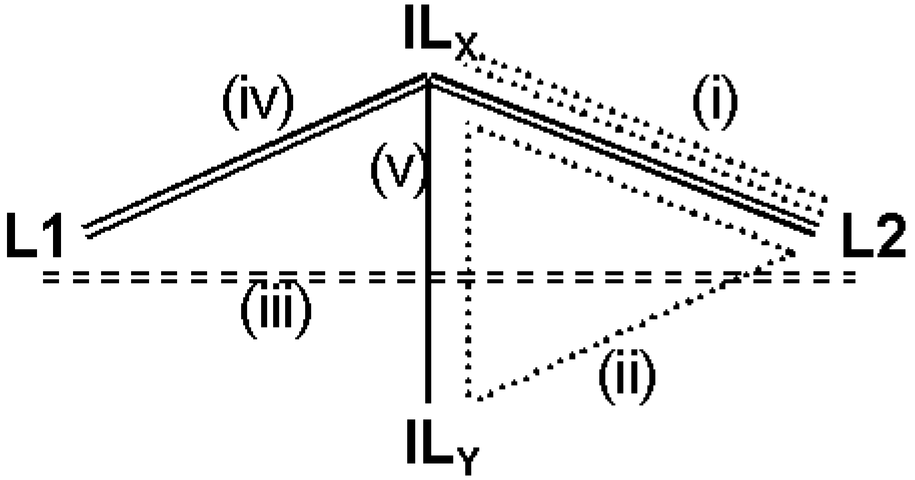
The use of learner corpora is often seen as one possible way to avoid the worst limitations of traditional EA.

1.1 Studying interlanguage with learner corpora

Learner corpora are a fairly new arrival on the corpus linguistic scene, but have quickly become one of the most important resources for studying interlanguage. Like other corpora, a learner corpus is “a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration” (McEnery and Wilson 2001: 32). A learner corpus is a collection of texts – written texts or transcribed spoken language – produced by language learners, and sampled so as to be representative of one or more combinations of situational and learner factors. This addresses the first limitation of EA mentioned in the preceding section; by design, learner corpus data is homogeneous.

The whole gamut of corpus linguistics methods and tools are applicable to learner corpora, too. Available for immediate application are such tools as concordancers and word (form) listing, sorting and searching utilities, as well as statistical processing on the word form level. Even with these fairly simple tools you can accomplish a lot, especially with ‘morphologically naive’ languages like English. For deeper linguistic analysis, learner corpora can be lemmatized, annotated for part-of-speech (POS) – or POS-tagged –and/or parsed to various degrees of complexity. Learner corpora can also be annotated for the errors found in them, which raises the intricate question of how errors are to be classified and corrected (Dagneaux et al. 1998). Utilizing methods from parallel corpus linguistics (Borin 2002a; Kilgarriff 2001),[1] learner corpora can be compared to each other or to corpora of texts produced by native speakers of the learners’ target language (L2) or their native language(s) (L1). Figure 1 illustrates some of the possibilities in this area.

Figure 1: *Learner corpora and SLA research*



In Figure 1, case (i) [the double dotted line] is the ‘classical’ mode of learner corpus use (and of traditional error analysis) – interlanguage analysis (IA).[2] Here, the interlanguage (IL), represented by the learner corpus, is compared to a representative native-speaker L2 corpus. Case (ii) [the dotted triangle] is an extension of (i), where different kinds of IL are contrasted to each other and to the L2 (called CIA – *contrastive interlanguage analysis* – by Granger 1996). The different ILs could be produced by learners with different native languages (as in most investigations based on ICLE; see Granger 1998 and 4.1 below) or by learners with different degrees of proficiency, or, finally, by the same learners at different times during their language learning process, i.e. a *longitudinal* comparison (Hammarberg 1999), which goes some way towards dealing with limitation 5 of EA (see above). Case (iii) [faint double dashed line] represents a methodological tool which at times has been important in SLA research, but not very much pursued in the context of learner corpora, namely *contrastive analysis* (CA), where native-speaker L1 and L2 are compared in order to find potential sources of interference.[3] Cases (i), (ii) and (iii) are quite general, and are meant to cover investigations on all linguistic levels. For pragmatic reasons, most such investigations have confined themselves to the level of lexis and such syntactic phenomena which are easily investigated through lexis. However, there is an increasing amount of work on (automatically) part-of-speech-tagged (POS-tagged) learner corpora (e.g., Aarts and Granger 1998; see 4.1 below), and even some investigations of parsed learner corpora (see Meunier 1989; Staerner 2001). The present paper addresses case (iv) [the double solid lines], which to the best of our knowledge has not been investigated earlier using learner corpora,[4] and in the future, we hope to be able to also look into case (v) [the double+single solid lines], the extension of case (iv) to more than one kind of IL.

2. Investigating syntactic interference in learner language

We now turn to our own investigation. In distinction to most other studies of learner language corpora, where the IL has been compared only to native L2 production, we add a comparison with the learners' L1. Arguably, this makes our study not only one of interlanguage in general, but of specific L1 *interference* as evidenced in IL, which is relevant, e.g., for the development of intelligent CALL applications, incorporating natural language processing components – our particular area of expertise – e.g. learner language grammars and learner models.

We investigated differences in the frequencies of POS sequences (or POS *n-grams*) between a corpus of native English on the one hand, and two corpora – one of Swedish advanced learner English and one of native Swedish – on the other hand, the hypothesis being that significant common differences would reflect L1 interference in the IL on the syntactic level, since the POS sequences arguably serve as a rough approximation of surface syntactic structure, at least in the case of languages where syntactic relations are largely signalled by constituent order (both English and Swedish are such languages). The differences found were of two kinds, reflecting *overuse* or *underuse* of particular POS sequences, common to Swedish advanced learner English and Swedish, as compared to native English. In what follows, we will refer to those IL traits that we focus on in our investigation as "IL+L1".[5]

2.1 The corpora and tagsets

For our investigation, we used the following three sets of corpus materials.

1. The learner corpus, the Uppsala Student English corpus (USE; Axelsson 2000; Axelsson and Berglund 2002), contains about 400,000 tokens (about 350,000 words);
2. The native English corpus was made up of the written language portion of the British National Corpus Sampler (BNCS; Burnard 1999), containing about 1.2 million tokens (roughly 1 million words);
3. The native Swedish corpus, the Stockholm Umeå Corpus (SUC; Ejerhed and Källgren 1997), contains roughly 1.2 million tokens (about 1 million words).

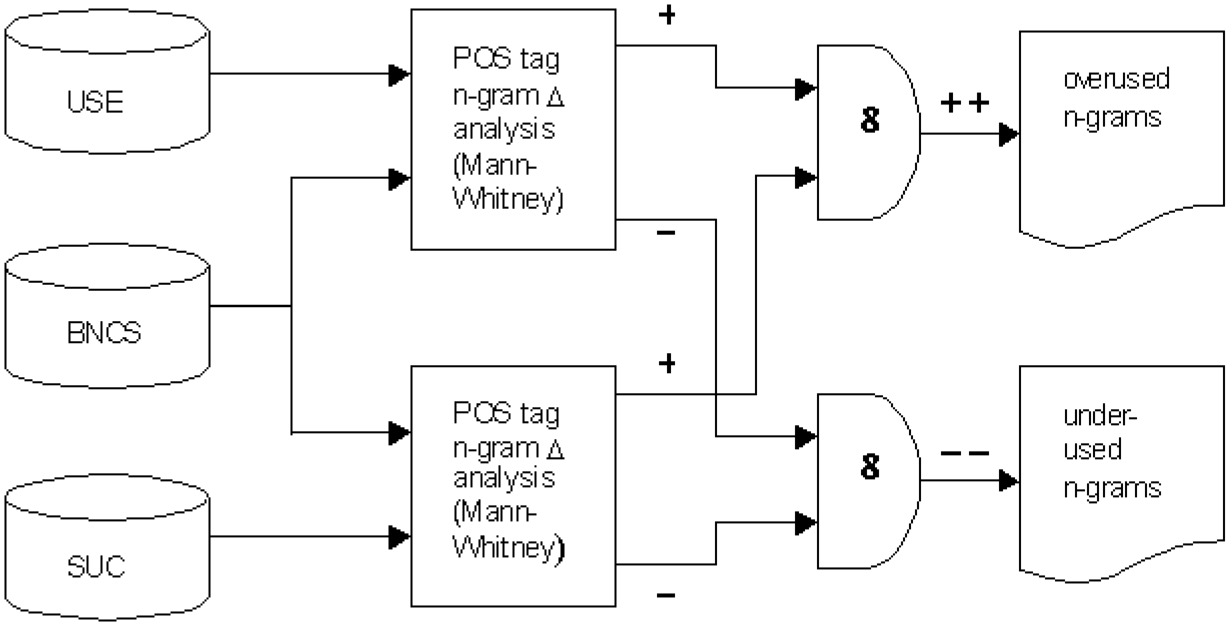
The BNCS and SUC corpora come in POS-tagged, manually corrected versions, which we have used without modification. The USE corpus was tagged by us with a Brill tagger trained on the BNC sampler, giving an estimated accuracy of 96.7 %. For the purposes of this investigation, both tagsets were reduced, the English set to 30 tags (from 148) and the Swedish to 37 tags (from 156). The reduced tagsets are listed and compared in the Appendix. The tagsets were reduced 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 (2002) experiment with a 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 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, 2002b).

2.2 Experiment setup

In Figure 2, the setup of the experiment is shown in overview. We used a similar procedure to that of our earlier investigation of translationese (Borin and Prütz 2001):[6]

1. First, we extracted all POS n-gram types (for n = 1 ... 4) and their frequencies from the three POS-tagged corpora;
2. From the n-gram lists we removed certain sequences, namely (a) those containing the tag NC (proper noun; we believe that a higher or lower relative incidence of proper nouns is not a distinguishing trait in learner language), (b) those with punctuation tags except for those containing exactly one full-stop tag, in the first or the last position,[7] and (c) those not appearing in all three corpora, either by necessity (because of differences between the English and Swedish tagsets) or by chance;
3. For each n-gram length, the incidence of the n-gram types in BNCS (representing native English) and USE (representing learner English) were compared, using the Mann-Whitney (or U) statistic (see Kilgarriff 2001 for a description and justification of the test for this kind of investigation), and instances of significant ($p \leq 0.05$, two-tailed) differences (overuse and underuse) were collected ("n-gram Δ analysis" in Figure 2);
4. BNCS and SUC (representing the learners' native language, i.e. Swedish) were compared in exactly the same way;
5. Finally, the n-gram types which showed significant overuse or significant underuse in both comparisons were extracted, symbolized by the "&" (logical AND) process in Figure 2.

Figure 2: *Experiment setup*


USE
BNCS
SUC
POS tag n-gram Δ analysis (Mann-Whitney)
POS tag n-gram Δ analysis (Mann-Whitney)
+
−
+
−
&
&
++
−−
overused n-grams
under-used n-grams

3. Results by the numbers

In this section, we give a general overview of our results, but defer discussion to section 4, where we compare our findings with those of other similar investigations. In Table 1, you will find the numbers, i.e. how many of each n-gram type occured in each corpus. We give both the actual and the theoretically expected figures. For unigrams, the expected figure is the cardinality of the tagset, of course, while the figure for the other n-grams is the actually occurring number of unigrams in the corpus in question raised to the corresponding power; thus, 29^3 (29 cubed) is the expected number of trigrams in the USE corpus. This simply illustrates the well-known fact that language has syntax, and is not in general freely combinatorial.

The longer the sequence, the smaller the fraction becomes that is actually used of all possible combinations. This is what makes it possible to let POS n-grams stand in for real syntactic analyses.

Table 1: *Actually occurring and expected n-gram types in the corpora*

corpus:	**USE**		**BNCS**		**SUC**	
	occurring	*(expected)*	*occurring*	*(expected)*	*occurring*	*(expected)*
unigrams	29	(30)	30	(30)	34	(37)
bigrams	663	(841)	807	(900)	1035	(1156)
trigrams	6526	(24389)	10800	(27000)	13616	(39304)
4-grams	31761	(707281)	60645	(810000)	72770	(1336336)

In Table 2, underuse and overuse are shown, found by the experimental procedure described in the previous section. The percentage figures shown in the table are calculated by dividing the underuse/overuse figures by the POS n-gram figures for the USE corpus, i.e., the percentage of significantly different (underused and overused) trigrams is calculated as (42+155)/6526 (≈ 0.03019, i.e. 3.0%). An interesting fact reflected by the figures in Table 2 is that there turned out to be more instances of overuse than of underuse for all n-gram lengths.

Table 2: *Underuse and overuse per n-gram length*

unigrams		**bigrams**		**trigrams**		**4-grams**	
underuse	*overuse*	*underuse*	*overuse*	*underuse*	*overuse*	*underuse*	*overuse*
1	3	11	36	42	155	91	171
3.4%	10.3%	1.6%	5.4%	0.6%	2.4%	0.3%	0.5%
= 13.7%		= 7.0%		= 3.0%		= 0.8%	

In section 3.1, we will discuss some representative cases of each n-gram type.

3.1 Distinctive IL+L1 n-grams

3.1.1 Unigrams

Among the unigrams, there was one instance of underuse, "K2" (past participle), while there were three overused parts-of-speech: "V" (finite verb), "R" (adverb), and "C" (conjunction). Possibly, this indicates a less complex sentence-level syntax in the IL+L1 than in native English, with more finite clauses joined by conjunctions, rather than non-finite subordinate clauses.[8] The adverbs could be a sign of a more lively, narrative style, and may possibly have nothing at all to do with the fact that these particular narratives happen to be in interlanguage (but see section 4.2).

3.1.2 Bigrams

Just as adverbs by themselves are overused in the USE IL+L1, so are a number of bigrams containing adverbs, e.g. "R C" (adverb–conjunction), "R R" (adverb–adverb), "R NN" (adverb–common noun), "R V" (adverb–finite verb), ". R" (sentence-initial adverb). Sentence-initial common nouns (". NN") are also overused, perhaps strengthening the impression that sentence syntax is simpler in IL+L1 than in native L2.

By way of illustration, we show some examples of the bigram "R R" from the USE corpus (the full tagset is used in this and in the other examples which follow below):

(1)

I/PPIS1 **also/RR recantly/RR** descovered/VVN that/CST
my/APPGE spelling/NN1 was/VBDZ rather/RG poor/JJ
so_that/CS is/VBZ someting/PN1 I/PPIS1 have/VH0 to/TO
work/VVI on/RP ./YSTP

(2)

He's/NP1 **far/RR away/RP** ./YSTP

(3)

So/RG naturally/RR ,/YCOM they/PPHS2 were/VBDR shocked/JJ
to/TO find/VVI complete/JJ wilderness/NN1 and/CC a/AT1
nature/NN1 so/RR unlike/II the/AT English/NN1 ./YSTP

Additionally, examples 4–6 in section 3.1.3 below also contain "R R".

All the most consistently underused bigrams have in common the POS tag "K2" (past participle): "K2 I" (past participle–preposition), "K2 R" (past participle–adverb), "NN K2" (common noun–past participle), "V K2" (finite verb–past participle). We give some examples of the "K2 R" bigram in section 3.1.4 below (examples 13–18), from which we see that the adverb (at least often) is the second component (the verb particle) of a phrasal (or particle) verb. Hence, the IL+L1 shows an underuse of either periphrastic tenses or non-finite clauses, or both, with phrasal verbs.[9]

3.1.3 Trigrams

Many of the overused trigrams contain adverbs: ". R R" (sentence-initial adverb–adverb; example 3), "R R NN" (adverb–adverb–common noun; examples 4–6). Other examples of overused trigrams are "VI I NN" (infinite verb–preposition–common noun; examples 10–12), "V I NN" (finite verb–preposition–common noun).

(4)

When/CS I/PPIS1 write/VV0 ,/YCOM I/PPIS1 can/VM spend/VVI
as/RG much/RR time/NNT1 as/CSA I/PPIS1 want/VV0 to/TO
make/VVI changes/NN2 and/CC corrections/NN2 ./YSTP

(5)

They/PPHS2 are/VBR trying/VVG to/TO imitate/VVI
their/APPGE action/NN1 heroes/NN2 and/CC not/XX **very/RG
seldom/RR** accidents/NN2 occur/VV0 ./YSTP

(6)

That_is/REX **however/RR far_from/RG** reality/NN1 ./YSTP

Among the underused trigrams we find many which contain adjectives: "A A NN" (adjective–adjective–common noun), "A NN K1" (adjective–common noun–present participle), "A NN K2" (adjective–common noun–past participle), "A NN NN" (adjective–common noun–common noun). Past participles appear among underused trigrams as well. Thus, we find "NN K2 R" (common noun–past participle–adverb) in addition to the already mentioned "A NN K2".

3.1.4 4-grams

Among overused 4-grams, there are a number involving conjunctions and prepositions, e.g.: ". C NN V" (sentence-initial conjunction–common noun–finite verb; examples 7–9), "C NN R V" (conjunction–common noun–adverb–finite verb), "VI I NN ." (sentence-final infinite verb–preposition–common noun; examples 10–12), "V I NN ." (sentence-final finite verb–preposition–common noun).

(7)

When/CS people/NN grew/VVD old/JJ they/PPHS2 were/VBDR
depending_on/II their/APPGE relatives'/JJ goodness/NN1
./YSTP

(8)

When/CS children/NN2 reach/VV0 a/AT1 certain/JJ age/NN1
,/YCOM they/PPHS2 tend/VV0 to/TO find/VVI these/DD2
violent/JJ films/NN2 very/RG cool/JJ and/CC exciting/JJ
./YSTP

(9)

Because/CS fact/NN1 is/VBZ that/CST New/JJ Lanark/NP1 was/VBDZ a/AT1 success/NN1 ,/YCOM a/AT1 large/JJ one/PN1 ./YSTP

(10)

I/PPIS1 have/VH0 always/RR found/VVN it/PPH1 amusing/JJ to/TO **write/VVI in/II English/NN1 ./YSTP**

(11)

We/PPIS2 need/VV0 to/TO teach/VVI them/PPHO2 how/RRQ to/TO defend/VVI themselves/PPX2 in/II today's/NN2 society/NN1 and/CC to/TO **turn/VVI away_from/II violence/NN1 ./YSTP**

(12)

Another/DD1 great/JJ fear/NN1 was/VBDZ that/CST wilderness/NN1 would/VM force/VVI civilised/JJ men/NN2 to/TO **act/VVI like/II savages/NN2 ./YSTP**

In the set of underused 4-grams, there are quite a few containing past participles, e.g.: “K2 R I A” (past participle–adverb–preposition–adjective), “K2 R I NN” (past participle–adverb–preposition–common noun; examples 13–15), “K2 R I P” (past participle–adverb–preposition–pronoun; examples 16–18), “NN V K2 R” (common noun–finite verb–past participle–adverb).

(13) Why/RRQ does/VDZ anyone/PN1 want/VVI to/TO see/VVI a/AT1 man/NN1 get/VV0 his/APPGE head/NN1 **chopped/VVN off/RP on/II television/NN1** ?/YQUE

(14) Tom/NP1 is/VBZ **blown/VVN up/RP with/IW dynamite/NN1** but/CCB is/VBZ still/RR alive/JJ ./YSTP

(15) You/PPY can/VM be/VBI **swept/VVN away/RP with/IW money/NN1** ,/YCOM towards/II materialistic/JJ values/NN2 ,/YCOM without/IW even/RR realizing/VVG it/PPH1 ./YSTP

(16) It/PPH1 is/VBZ essential/JJ to/II all/DB infant/NN1 mammals/NN2 to/TO be/VBI taken/VVN care/NN1 of/IO ,/YCOM and/CC to/TO be/VBI **brought/VVN up/RP by/II someone/PN1** who/PNQS knows/VVZ the/AT difficulties/NN2 of/IO life/NN1 ./YSTP

(17) However/RR ,/YCOM the/AT Chief's/NN2 images/NN2 of/IO machines/NN2 are/VBR not/XX only/RR similes/VVZ ,/YCOM he/PPHS1 also/RR suffers/VVZ delusions/NN2 which/DDQ make/VV0 him/PPHO1 think/VVI that/CST there/EX are/VBR actual/JJ machines/NN2 **installed/VVN everywhere/RL around/II him/PPHO1** ,/YCOM controlling/VVG him/PPHO1 ./YSTP

(18) I/PPIS1 know/VV0 that/CST woman/NN1 is/VBZ naturally/RR and/CC necessarily/RR weak/JJ in_comparison_with/II man/NN1 ;/YSCOL and/CC that/CST her/APPGE lot/NN1 has/VHZ been/VBN **appointed/VVN thus/RR by/II Him/PPHO1** who/PNQS alone/JJ knows/VVZ what/DDQ is/VBZ best/JJT for/IF us/PPIO2 ./YSTP

4. Comparisons with similar previous work

In this section we compare our results in more detail to other relevant work. The only similar investigation of learner language that we know of is that made by Aarts and Granger (1998). Their work is methodologically similar to our approach, and therefore a fairly detailed comparison or our findings with theirs seems warranted. Section 4.1 is devoted to such a comparison. Further, it seems reasonable to assume that there should be common traits in translated language (*translationese*; Gellerstam 1985, 1996) and (advanced) learner language, and in section 4.2, we compare our results here to those obtained in our earlier investigation of translationese.

4.1 Aarts and Granger 1998

Aarts and Granger (1998; henceforth A&G) compared POS trigram frequencies in three learner corpus materials, the Dutch, Finnish and French components of ICLE, with comparable material produced by native speakers of English, i.e. the LOCNESS (LOuvain Corpus of Native English eSSays) corpus. Their investigation was thus an instance of corpus-based CIA (see above), and did not involve the native languages of the learners, other than indirectly, through the comparison between the three learner corpus materials.

A&G produced POS trigram frequency lists from all four corpus materials (each about 150.000 words in length). Like in our investigation, they worked with a reduced version of the tagset they used for tagging the corpora (the TOSCA-ICE tagset with 270 tags, which were reduced to 19). They then investigate their trigram lists in a number of ways:

1. They calculate significant differences (underuse and overuse in relation to LOCNESS) in the rank orderings of the lists, using the χ^2 test;
2. They investigate the differences common to the three ICLE components in relation to LOCNESS (the "cross-linguistic invariants"; about 7% of the trigrams),
3. and differences unique to one learner variety ("L1-specific patterns"; about 20–25% of the trigrams, depending on the L1), where only the French variety is discussed in any detail by A&G (see above).

We now proceed to a more detailed comparison between the findings of A&G and our own results (B&P in what follows). We should keep some things in mind, though. First of all, A&G actually make a different investigation. They investigate over- and underuse of POS trigrams in a learner corpus, compared to a native speaker corpus. Our investigation started out in the same way, but additionally, we remove all POS n-grams which do not differ in the same way between the native L2 corpus and a corpus of native L1, i.e. the native language of the learners. Thus, the POS n-grams that remain in our case should exclude A&G's "cross-linguistic invariants", if indeed their "L1-specific patterns" reflect transfer from the learners' native language. A&G use a smaller tagset (which reflects a partly different linguistic classification) than we do. Also, we have used a different statistical test for significance testing. These circumstances conspire to make comparisons between our investigations not entirely straightforward, and could easily account for the differences in the numbers that the two investigations arrive at (we come nowhere near the at least 20% L1-specific trigrams found by A&G; see Table 2, above). If our respective studies really investigate the same thing, we would make the following two predictions.

1. There could—but need not—be partial overlap between the "L1-specific patterns" A&G found and those that we have uncovered. The overlap should in that case be larger, the closer the L1 in question is to Swedish, i.e. A&G's Dutch ICLE material should show most overlap with our results. Unfortunately, A&G present concrete results only for French L1-specific patterns, which show practically no overlap with our patterns, as expected;
2. We would also predict that those POS trigrams that A&G found to be over- or underused in all the three subcorpora they investigated – the "cross-linguistic invariants" –, should not appear in our material. By and large, this prediction holds, i.e. most of the patterns that A&G find as significantly different in the same way in all the three L1-specific subcorpora, are indeed not present in our set of significantly differently distributed POS n-grams. The only possible exceptions to this are the sentence-initial pattern shown in Table 3, where the picture is not as clear as in other cases. Although there are so few n-grams that no firm conclusions can be drawn from them, it still seems that there is a

difference between those patterns where A&G found overuse and the ones that are underused according to their results. A&G tags should be fairly self-explanatory (except perhaps "#", sentence break), and B&P tags are explained in the Appendix. Differences are noted using "+" (overuse), "–" (underuse), and "≅" (no significant difference).

Table 3: *Comparison with language-invariant sentence-initial patterns found by A&G (based on Section 4.2, Arts and Granger 1998: 137)*

A&G POS sequence	= B&P POS sequence	A&G	B&P
overused			
# # CONNEC	. C	+	+
# # ADV	. R	+	+
# # PRON	. P	+	≅
underused			
# # N	. NN	–	+
# CONJ N	. C NN	–	≅
# PREP Ving	. I K1	–	≅

4.2 Borin and Prütz 2001

Intuitively, translated language (translationese; see above) and IL ought to have features in common: "Both are situated somewhere between L1 and L2 and are likely to contain examples of transfer." (Granger 1996: 48). Thus, it is of value to compare the results of the present investigation to an earlier similar investigation of translationese (Borin and Prütz 2001), where we looked at newstext translated from Swedish to English, using an almost identical experimental procedure to the one presented here. The differences were as follows.

1. Different corpora were used, of course: (a) The English translation and (b) Swedish original versions of a Swedish news periodical for immigrants, the "press, reportage" parts of the (c) Flob and (d) Frown English corpora;
2. In addition to the 1- – 4-grams investigated in IL+L1, we also investigated 5-grams in our translationese study;
3. The initial selection of distinct n-grams was different, and based on an absolute difference in rank in the corpora, rather than on a statistical test. The same set of n-grams as in the present investigation were then removed from consideration (i.e., those containing proper names and certain kinds of punctuation, and those not occurring in all the compared corpora; see above);
4. The statistical test was applied only to the results of the initial selection, resulting in the removal of a number of n-grams. However, we do not know if the initial selection has excluded some n-grams which would have been singled out as significantly different by the statistical test.

If we take as our hypothesis that there should be a fair amount of overlap between the two sets of distinct n-grams, or perhaps even that the n-grams found to be characteristic of translationese should be a subset of those characteristic of learner language, we have to admit that the hypothesis was soundly falsified.

What we found was that there were a considerably larger number of significant differences characteristic of learner language than of translationese (506 2- – 4-grams in IL+L1 vs. 41 in translationese), except in the case of unigrams, where IL+L1 had 4, against 6 in translationese. On the other hand, there is almost no overlap – let alone inclusion – between the two sets of n-grams. There are two shared bigrams (". R" and "C VI", both overused), one shared trigram (". I P", overused), and no shared unigrams or 4-grams.[10] The one similarity that we did find was a somewhat similar situation with regard to overuse and underuse. There are more overused than underused bigrams and trigrams both in IL+L1 and translationese, while they differ with respect to 4-grams, where translationese displayed more underuse than overuse.

In conclusion: While our results perhaps do not invalidate the intuition that IL and translationese "are situated somewhere between L1 and L2 and are likely to contain examples of transfer" (see above), it certainly seems that they are situated in quite different locations in the region between L1 and L2 (but see the next section). More research is clearly needed here.

5. Discussion and conclusion

In this section, we would like to discuss some general issues which bear on the interpretation of our results and on the comparisons we have made of these results with the findings of other similar investigations:

1. Representativeness of the English "standard". We have used (the written part) of BNCS as the L2 standard. Perhaps we should instead have used a native students' essay corpus such as LOCNESS (like Aarts and Granger 1998), or perhaps even a corpus of spoken English, acknowledging the fact that the written English of Swedish learners is held to be influenced by colloquial spoken English (see Hägglund 2001);
2. Representativeness of the Swedish "standard". In the same way, we could question whether SUC really faithfully represents the learners' "point of departure", the form of Swedish most likely to influence their IL English. Perhaps here, too, a corpus of spoken Swedish would serve better (see Allwood 1999), or possibly a corpus of Swedish student compositions;
3. What do the "L1-specific" trigrams found by Aarts and Granger (1998) reflect? Our hypothesis – which informed the way we set up our experiment, described in section 2 above – was that they represent transfer, i.e., underuse and overuse of an n-gram type in IL reflect relatively lower and higher incidence, respectively, of the same n-gram type in the L1. Only if this hypothesis holds are our results comparable with those of Aarts and Granger. If underuse or overuse in IL is due to something else, then obviously we cannot compare our results. E.g., underuse in the IL could be due to *avoidance* of an L1 structure, in which case it should be correlated to a higher incidence in the L1 or no significant difference;
4. There is an estimated tagging error rate of a bit more than 3% in the USE corpus (see section 2.1). If the errors made by the tagger are not random, there will be a bias in the results of our investigation;
5. POS tag sequences are of course not syntactic units; they merely give better clues to syntax than word-level investigations are able to provide, so that the picture we get of learner (and native speaker) language syntax is distorted and needs careful interpretation to be usable.

In conclusion, we would like to say that we think that our investigation confirms the observation made by Aarts and Granger (1998) and Borin and Prütz (2001) that a contrastive investigation of POS-tagged corpora can yield valuable linguistic insights about the differences (and similarities) among the investigated language varieties. At the same time, much remains to be done regarding matters of methodology; among others, the issues mentioned above need to be addressed.

In the future, we would like to look into the issue of L1 and L2 corpus representativeness. We would also like to extend and refine our investigation of L1 interference in learner language syntax in various ways, notably by the use of robust parsing (Abney 1996), which would enable us to look at syntax directly, to investigate e.g. which syntactic constituents and functions are most indicative of learner language.

Acknowledgements

We would like to thank the volume editors for their careful reading and commenting of (the previous version of) this article. The research presented here was funded by the following sources: an Uppsala University, Faculty of Languages *reservfonden* grant; Vinnova through the CrossCheck project; The Knut and Alice Wallenberg Foundation through the Digital Resources in the Humanities project, part of the Wallenberg Global Learning Network initiative.

Notes

[1] We use the term "parallel corpus linguistics" to subsume both work with *parallel corpora* – i.e., original texts in one language and their translations into another language or other languages – and work with *comparable corpora*, i.e., original texts in two or more languages which are similar as to genre, topic, style, etc. At least in the language technology oriented research tradition, there are interesting commonalities between the two kinds of work. (see Borin 2002a), e.g. in the use of distributional regularities – as revealed by statistical or information-theoretical measures – for automatically discovering translation equivalents in both kinds of corpora. Work such as that presented here, dealing

with comparisons among learner IL corpora and original L1 and L2 corpora, is most similar to work on comparable corpora, of course.

[2] But using a learner corpus and (computational) corpus linguistics tools, we can do much more than in traditional EA. Perhaps the major advantage is that we can investigate *patterns of deviant usage* – e.g., instances of overuse and underuse – rather than just instances of clear errors. Even in the latter case, we can generalize over the normal linguistic contexts (on many linguistic levels, to boot) of particular errors fairly easily using corpus linguistics tools, something which in general was not feasible in traditional EA. This takes care of limitations 3 and 4 of EA mentioned above.

[3] In corpus linguistics – at least if we are talking about the more interesting case, namely the development of *automatic* methods for making linguistically relevant comparisons between texts –, the closest thing to CA is the work on parallel and comparable corpora aimed mainly at extracting translation equivalents for machine translation or cross-language information retrieval systems (see, e.g., Borin 2002a). These methods, although at present used almost exclusively for language technology purposes, could in principle be used for a more traditionally linguistically-oriented "contrastive corpus linguistics" as well, as we have argued elsewhere (e.g. Borin 2001; cf. Granger 1996), complementing the largely manual modes of investigation used in present-day corpus-based contrastive linguistic research.

[4] At least not in the way that we propose to do it. Although it shares some traits with Granger's (1996: 46ff) proposed "integrated CA/CIA contrastive model [which] involves constant to-ing and fro-ing between CA and CIA", we believe that our method provides for a tighter coupling between all the involved language varieties; there is no difference (indeed, there *should* be no difference) between CA and IA with our way of doing things.

[5] Note that our method of investigation by design is unsuited for finding errors, since we count as instances of overuse only such items that actually appear in the native L2 corpus, i.e., it is not counted as an instance of overuse if a construction appears in the L1 and IL corpora but not in the L2 corpus (even though the difference in itself may be statistically significant). Concretely, this is achieved by taking the L2 corpus – i.e., the BNCS in our case –as the basis for all comparisons; see further 2.2 below.

[6] There were some small differences, which we will return to below, when we compare the results of the two investigations.

[7] The motivation for this is possibly less well-founded than in the case of proper nouns, but let us simply say that we wish to limit ourselves, at least for the time being, to looking at clause-internal syntax imperfectly mirrored in the POS tag sequences found in a text. Of course, at the same time we eliminate e.g. commas functioning as coordination conjunctions, 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 interference. 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.

[8] English has more possibilities for non-finite clausal subordination than Swedish, which may be relevant here. It seemed that the results of our earlier translationese investigation reflected this circumstance (Borin and Prütz 2001: 36). Granger (1997) finds a similar underuse of non-finite subordinate clauses in non-native written academic English as compared to that of native writers.

[9] Here, it would be good to compare our results with Hägglund's (2001) lexical investigation of phrasal verbs in the Swedish component of ICLE, compared to LOCNESS. For the time being, this will have to remain a matter for future investigation, however.

[10] Although it is an intriguing fact that one of the findings of our translationese study was significantly more adverbs in Swedish than in all the English materials, and that the English translated from Swedish had more – but not significantly more – than either of the other two English materials (see section 3.1.1).

References

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

Abney, S. 1996. "Part-of-speech tagging and partial parsing. In *Corpus-Based Methods in Language and Speech*, K. Church, S. Young and G. Bloothooft (eds). Dordrecht: Kluwer.

Allwood, J. 1999. "The Swedish spoken language corpus at Göteborg University". In Fonetik 99: Proceedings from the 12th Swedish phonetics conference. [Gothenburg papers in theoretical linguistics 81]. Department of Linguistics, Göteborg University.

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. 2002. "The Uppsala Student English Corpus (USE): a multi-faceted resource for research and course development". In *Parallel Corpora, Parallel Worlds*, L. Borin (ed), 79–90. Amsterdam: Rodopi.

Borin, L. 2000. “Something borrowed, something blue: Rule-based combination of POS taggers”. *Second International Conference on Language Resources and Evaluation. Proceedings, Volume I*, 21–26. Athens: ELRA.

Borin, L. 2001. “Att undersöka språkmöten med datorn”. In *Språkets gränser och gränslöshet. Då tankar, tal och traditioner möts. Humanistdagarna vid Uppsala universitet 2001*, A. Saxena (ed), 45–56. Uppsala: Uppsala University.

Borin, L. 2002a. “… and never the twain shall meet?”. In *Parallel Corpora, Parallel Worlds*, L. Borin (ed), 1–43. Amsterdam: Rodopi.

Borin, L. 2002b. “Alignment and tagging”. In *Parallel Corpora, Parallel Worlds*, L. Borin (ed), 207–218. Amsterdam: Rodopi.

Borin, L. and Prütz, K. 2001. “Through a glass darkly: Part of speech distribution in original and translated text”. In *Computational Linguistics in the Netherlands 2000*, W. Daelemans, K. Sima’an, J. Veenstra and J. Zavrel (eds), 30–44. Amsterdam: Rodopi.

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].

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

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

Ellis, R. 1985. *Understanding Second Language Acquisition*. Oxford: Oxford University Press.

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

Gellerstam, M. 1996. “Translations as a source for cross-linguistic studies”. In *Languages in Contrast. Papers from a Symposium on Text-Based Cross-Linguistic Studies. Lund 4–5 March 1994*, K. Aijmer, B. Altenberg and M. Johansson (eds), 53–62. Lund: Lund University Press.

Granger, S. 1996. “From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora”. In *Languages in Contrast. Papers from a Symposium on Text-Based Cross-Linguistic Studies. Lund 4–5 March 1994*, K. Aijmer, B. Altenberg and M. Johansson (eds), 37–51. Lund: Lund University Press.

Granger, S. 1997. “On identifying the syntactic and discourse features of participle clauses in academic English: native and non-native writers compared”. In *Studies in English Language and Teaching*, J. Aarts, I. de Mönnink & H. Wekker (eds.), 185–198. Amsterdam: Rodopi.

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

Hägglund, M. 2001. “Do Swedish advanced learners use spoken language when they write in English?”. *Moderna språk* 95 (1): 2–8.

Hammarberg, B. 1999. “Manual of the ASU Corpus – a longitudinal text corpus of adult learner Swedish with a corresponding part from native Swedes”. Stockholm University, Department of Linguistics.

Kilgarriff, A. 2001. “Comparing corpora”. *International Journal of Corpus Linguistics* 6 (1): 1–37.

McEnery, T. and Wilson, A. 2001. *Corpus Linguistics*. 2nd edition. Edinburgh: Edinburgh University Press.

Meunier, F. 1998. “Computer tools for the analysis of learner corpora”. In *Learner English on Computer*, S. Granger (ed), 19–37. London: Longman.

Prütz, K. 2002. “Part-of-speech tagging for Swedish”. In *Parallel Corpora, Parallel Worlds*, L. Borin (ed), 201–206. Amsterdam: Rodopi.

Staerner, A. 2001. “Datorstödd språkgranskning som ett stöd för andraspråksinlärning” [Computerized language checking as support for second language learning]. MA Thesis in Computational Linguistics, Department of Linguistics, Uppsala University. Retrievable via <http://ww.ling.uu.se/>.

Appendix: Reduced Swedish and English tagsets

Table A1: *Reduced Swedish (SV-R) and English (EN-R) tagsets*

	SV-R	**EN-R**		**description**	**examples**
1	–	–	1	dash	–
2	!	!	2	exclamation mark	!
3	"	"	3	quotes	”
4	(	(	4	left bracket	(
5	)	)	5	right bracket	)
6	,	,	6	comma	,
7	.	.	7	full-stop	.
		...	8	ellipsis	...
8	:	:	9	colon	:
9	;	;	10	semicolon	;
10	?	?	11	question mark	?
		$	12	genitive clitic	’s
11	A	A	13	adjective	röd, red
12	C	C	14	conjunction	och, that
13	E	E	15	infinitive mark	att, to
14	F			numeric expression	16
15	G			abbreviation	d.v.s.
16	I	I	16	preposition	på, on
17	K1	K1	17	present participle	seende, eating
18	K2	K2	18	past participle	sedd, eaten
19	L			compound part	hög-
20	M	M	19	numeral	två, two
21	NC	NC	20	proper noun	Eva, Evelyn
22	NC$			proper noun, genitive	Åsas
23	NN	NN	21	noun	häst, goat
24	NN$			noun, genitive	tjuvs
25	O	O	22	interjection	bu, um
26	P	P	23	pronoun	vi, we
27	P$	P$	24	pronoun, poss. or gen.	vår, our
28	Q			pronoun, relative	som
29	R	R	25	adverb	fort, fast
30	S	S	26	symbol or letter	G
31	T	T	27	determiner	en, the
32	V	V	28	verb, finite	såg, ate
33	VI	VI	29	verb, infinitive	se, eat
34	VK			verb, subjunctive	såge
35	VS			verb, supine	sett
36	X	X	30	unknown or foreign word	
37	ERROR			(tagged at all only in SUC)	