Literary and Linguistic Computing Advance Access published October 11, 2010

Automatically Extracting Typical Syntactic Differences from Corpora

Wybo Wiersma
King's College London

John Nerbonne
University of Groningen

Timo Lauttamus
University of Oulu

Abstract

We develop an aggregate measure of syntactic difference for automatically finding common syntactic differences between collections of text. With the use of this measure, it is possible to mine for differences between, for example, the English of learners and natives, or between related dialects. If formulated in advance, hypotheses can also be tested for statistical significance. It enables us to find not only absence or presence, but also under- and overuse of specific constructs. We have applied our measure to the English of Finnish immigrants in Australia to look for traces of Finnish grammar in their English. The outcomes of this detection process were analysed and found to be insightful. A report is included in this article. Besides explaining our method, we also go into the theory behind it, including permutation statistics, and the custom normalizations required for applying these tests to syntactical data. We also explain how to use the software we developed to apply this method to new corpora, and give some suggestions for further research.

Correspondence:
Wybo Wiersma,
St. Cross Annexe,
10 St. Cross Road,
Oxford OX1 3TU, UK.
E-mail:
mail@wybowiersma.net

1 Introduction

Languages are always changing and never homogenous or completely isolated. For example, language contact is a common phenomenon, and one which may even be growing due to the increased mobility of recent years. There are also differences in language usage between various subcultures in society, whether or not under the influence of education and media. And lastly, there are of course differences between regional dialects, which might also be changing under the influence of the above-mentioned, and many other, factors, including their own complex internal dynamics.

But as these rich fields for investigation are being explored, we nonetheless still seem to lack ways of assaying the aggregate differences in language usage between various groups. For example, most of the cross-linguistic research into second language acquisition (SLA) has so far focused on examining typical second-language learners' errors, such as absence of the copula, absence of prepositions, different (deviant) uses of articles, loss of inflectional endings, and deviant word order. These are often examined for

doi:10.1093/llc/fqq017

evidence of interference (in learning) or potential substrate influence in contact situations (Odlin, 1989, 1990, 2006, 2009). As Weinreich famously noted:

> No easy way of measuring or characterizing the total impact of one language on another in the speech of bilinguals has been, or probably can be devised. The only possible procedure is to describe the various forms of interference and to tabulate their frequency. (Weinreich, 1968, p. 63)

This article proposes, explains, and tests a computational technique for measuring the aggregate degree of syntactic difference between two varieties of language. With it we attempt to measure the 'total impact' in Weinreich's sense, albeit with respect to a single linguistic level, syntax. It may make the data of many linguistic, sociolinguistic, and dialectological studies amenable to the more powerful statistical analysis currently reserved for numerical data.

Naturally, researchers want more than a measure which simply assigns a numerical value to the difference between two syntactic varieties. We also want to know how significant the difference is, and we want to be able to identify the sources of the difference, not only in order to win confidence in the measure, but also to answer linguistic questions, such as those about the relative stability/volatility of syntactic structures. The technique presented not only offers a significance value for the aggregate difference, but also allows one to pinpoint the syntactic differences responsible for it. Strictly speaking, such significant values are of course only valid if hypotheses are formulated in advance. Lauttamus *et al.* (2007) present some linguistic results of the technique when applied to a corpus of English conversation with emigrants from Finland to Australia. The present article focuses on the technical and mathematical basis of the technique.

In the second and third sections, we introduce, explain, and discuss our method at an intuitive level. In the fourth, we go into some practical results it has produced as applied to the English of Finnish immigrants in Australia. The fifth will go deeper into the statistical theory behind it, and contains a full, mathematical description of the method. Then the sixth and seventh sections contain an explanation of how to use the software we produced for doing this research, and some suggestions for possible future research. We then conclude and provide some acknowledgements.

1.1 Related work

Thomason and Kaufmann (1988) and van Coetsem (1988) noted, nearly simultaneously, that the most radical (structural) effects in language contact situations are to be found in the language of 'switchers', i.e. in the language used as a second or later language. In line with this, we looked at the English of immigrants in our example research.

Poplack and Sankoff (1984) introduced techniques for studying lexical borrowing and its phonological effects, and Poplack *et al.* (1988) went on to exploit these advances in order to investigate the social conditions in which contact effects flourish best.

We follow Aarts and Granger (1998) most closely, who suggest focusing on tag sequences in learner corpora, just as we do. We shall add to their suggestion a means of measuring the aggregate difference between two varieties, and show how we can test whether that difference is statistically significant.

Nathan Sanders (2007) has, in the meantime, extended our method to use parse tree leaf-path ancestors of Sampson (2000) instead of *n*-grams. These are the tags along the routes from the root tag of the parse tree up to and not including each word of the sentence. He used it on the British part of the *ICE* (*International Corpus of English*) corpus, which is already fully parsed. In this article, however, we only report on the method using *n*-grams, but we find his extensions promising.

Related work on classifying syntaxes at an aggregate level may be found in the authorship recognition literature. Baayen *et al.* (1996) work with full parses on an authorship recognition task, while Hirst and Feiguina (2007) apply partial parsing in a similar study, obtaining results that allow them to distinguish a notoriously difficult author pair, the Brontë sisters. Also, they establish that their technique can work for even short texts (500 words and fewer), which could be an enormous advantage in applications of methods such as the one presented here or even a combination of our methods.

2 Methods

The fundamental idea of the proposed method is to tag the corpus to be investigated syntactically, to create frequency vectors of *n*-grams (trigrams for example) of POS (part-of-speech) tags, and then to compare and analyse these using a permutation test. This then results in both a general measure of difference and a list with the POS-*n*-grams that are most responsible for the difference.

In five steps the method proceeds:

- POS-tag two collections of comparable material;
- take *n*-grams (1- to 5-grams) of POS-tags from it
- compare their relative frequencies using a permutation test;
- sort the significant POS-*n*-grams by extent of difference;
- analyse the results.

Now we will describe these steps in more detail, starting with the tagging of two comparable collections of the text.

2.1 Tagging two collections of comparable material

We start with two collections of comparable material. For this, you can think of two sets of interviews with people from different dialect areas, or essays from two different grades and other similar pairs of samples. In our example research into the English of Finnish immigrants in Australia, we used interviews divided in two generations of arrival in Australia. One group was aged 16 years or younger when they disembarked, and the other was 17 years or older. The interviews come from the Finnish Australian English Corpus by Greg Watson (1996), and are reduced to the parts that were free conversation, leaving a remaining total of 305,000 words. The FAEC corpus is available from: http://wybowiersma.net/pub/fiauimenre/faec.tgz.

Then the material is POS-tagged, or in other words, lexical categories are assigned to all words in all the sentences. There are many POS-taggers, both statistical and rule-based taggers;[1] but we use Thorsten Brants' *Trigrams 'n Tags* (*TnT*) tagger, a hidden Markov model tagger that has performed at state-of-the-art levels in organized comparisons, achieving a precision of 96.7% correct tag assignments on the material of the *Penn Treebank* corpus (Brants, 2000).

As our chosen tagger is a statistical tagger, we also need a tagset and a corpus to train it on. For this, we choose the British part of the *ICE* corpus (Nelson *et al.*, 2002). This corpus is fully tagged and checked by hand, so it forms a sound basis. It uses the *TOSCA-ICE* tagset, which is a linguistically sensitive set, designed by linguists (not computer scientists) and consisting of 270 POS tags (Garside *et al.*, 1997)

2.2 Taking *n*-grams of POS-tags

Then, in order to be able to look at syntax instead of just single POS-tags, the tags are collected into *n*-grams, i.e. sequences of POS tags as they occur in corpora. For a sentence such as 'We'll have a roast leg of lamb tomorrow (...)' (extracted from our data), the tagger assigns the following POS labels:

We	'll	have
PRON(pers,plu)	AUX(modal,pres,encl)	V(montr,infin)
a	roast	leg
ART(indef)	N(com,sing)	N(com,sing)
of	lamb	tomorrow
PREP(ge)	N(com,sing)	ADV(ge)

These are then collected into *n*-grams. Trigrams are as follows: PRON(pers, plu)-AUX(modal, pres, encl)-V(montr, infin), . . . , ART(indef)-N(com, sing)-N(com, sing), . . . , PREP(ge)-N(com, sing)-ADV(ge). . .

In our research into the English of Finnish immigrants to Australia we used POS trigrams, and collected about 47,000 different kinds of trigrams from our corpus. For optimization reasons, and for a reason that we will come back to in Section 5.3.2, we removed all trigrams that occurred five times or less, leaving us with a remaining total of 8,300 POS-trigram-types.

2.3 Compare frequencies using a permutation test

The next step consists of counting how frequently each of the POS-*n*-grams (the 8,300 trigrams in our case) occurs in both of the datasets, or sub-corpora. These counts result in two vectors, or in more familiar phrasing, in two table rows in a $2 \times 8{,}300$ element table, with for each *n*-gram a column of two cells, containing frequency counts, one for each group. These form the input for our statistical test. Although there is no convenient test in classical statistics that we can apply to check whether the differences between vectors containing 8,300 elements are statistically significant, we may fortunately turn to permutation tests with a Monte Carlo technique in this situation (Good, 1995). We will explain it in detail (with all formulas) in Section 5, but the fundamental idea in a permutation test is very simple.

We first measure the difference between two sets of data in some convenient fashion, obtaining a degree of difference, lets call this the base case. We then extract two sets at random from the total of all data pooled together, which become our new sets, and we calculate the difference between these two in the same fashion. We then look if the difference is the same or bigger than in the base case, and if it is, we take note of this. We repeat this process with the random sets 10,000 times, taking different random sets every time, and in the end we sum the total number of times the difference was at least as extreme as in the base case. This value is then divided by the ten thousand times we tried, and the outcome of that is—in standard hypothesis testing terms—our *p*-value. This represents how many times in 10,000 (standardized to one) a difference as large as that found in the base case would have occurred by pure chance.

But before a statistical difference can be determined, a measure for the difference has to be chosen, and we have to decide what elements to permutate. For various statistical reasons explained in Section 5.1, we permutated speakers, using what we call between-subjects normalization. We also normalized for frequency using a between-types normalization (see Section 5.3 for these). As our measure we developed *RSquare*, which takes the square of the difference between each normalized *n*-gram count for the two groups, more on this in Section 5.4.

Thus using two normalizations and a suitable measure we permutated authors instead of sentences or *n*-grams and the test provides us with a *p*-value for the overall difference between the two sets of data.

2.4 Sort POS-*n*-grams by extent of difference

In addition to testing the aggregate difference between the two sub-corpora, we also wanted to be able to extract the POS-*n*-grams that were most responsible for the difference. Finding individual POS-*n*-grams allowed us to find the linguistic sources of the difference, and to understand these better. In addition, it also enabled us to test the method.

We get a list of responsible POS-*n*-grams by looking at the vectors from the base case again, and applying a permutation test to all the individual POS-*n*-grams. In other words, we do 8,300 permutation tests, one for each *n*-gram. So for each individual *n*-gram, the *RSquare* value of the base case is kept, and then in each permutation the *RSquare* value for the same *n*-gram is compared to it for being as large or larger, counting these, and using them to arrive at *p*-values, as in the original permutation test. For practical reasons, we did these tests all at the same time, by keeping track of and comparing all the *n*-gram *RSquare* values, besides the aggregate *RSquare* value. It might look like there is issue with so-called family-wise errors, because of the 8,300 tests, of which 5% (about 415) would get a $p < 0.05$ by pure chance. We cover this by requiring the aggregate difference to be significant at a *p*-level of 0.05 first (as an omnibus test, for the possible truth of the overall null-hypothesis; there being no overall difference). This protects us from complete null hypothesis family-wise errors (Westfall and Young, 1993). More on this, and possible improvements, in Section 6.

Once we have this list of significant POS-*n*-grams, we sift them into the groups for which they are typical. We do this for each *n*-gram by comparing the value found in the base case for that *n*-gram with the expected value for it based on the group size. If the groups are of the same size, the expected values are the same for both groups (half the corpus-wide count for the *n*-gram) and it is simply a case of looking at which of the base case values is larger, but if group sizes differ, we need to use the expected values based on group size (see the third normalization, Section 5.3.3).

So we have two lists now, one for each group. Then for each group we sort the *n*-grams in its list by how characteristic the *n*-gram type is for that group. The extent of typicality can be determined both relatively, normalized for frequency, and absolutely, but more on these options in Sections 5.3.2 and 5.3.1. What is important here is that they are sorted in this step.

As noted we only select and sort the significant POS-*n*-grams. The requirement of significance filters out any *n*-grams which occur only or mostly in one group, but for which this is probably due to chance. For example, if we simply selected *n*-grams for which more than, say, 80% of occurrences fall in one group, then we would also select *n*-grams occurring only once, purely by chance, in one of the groups (these would be 100% characteristic, but never significant).

At the end of this step, we thus have two lists of typical, significant POS-*n*-grams which can be analysed for linguistic causes of the aggregate difference between the groups. Moreover, we sort them to bring the most characteristic *n*-grams to the top, so that we see the most relevant data first. We may then analyse the data only up to a specific cut-off point, such as for example the top 200 of each list.

2.5 Analyse the results

The last step consists of analysing the data. This comes down to putting the top X *n*-grams in context and attempting to interpret them. This obviously requires intimate knowledge of the languages and/or dialects under comparison. To make things easier, besides the toolset for the method, we also developed some tools to help the analysis, e.g. a program to find examples from the corpus for given POS-*n*-grams.[2]

Besides tools and linguistic skills there are two important facts to keep in mind while analysing the results. The first is that the method will find differences only between the two sets provided for comparison. For example, if a certain POS-*n*-gram is over- or under-used in both groups, relative to the general population, then the method will not find it as being characteristic for both groups. So one has to be careful not to make claims that involve a comparison to the broader population of native speakers.

A second thing to note is that when a certain (set of) POS-*n*-grams is significantly characteristic for a group, this does not mean that ones explanation is correct. We do not claim that POS-*n*-grams play an explanatory role in the account of language contact, only that they are indicative of language contact effects. There can be more than one explanation for the over- or under-use of POS-*n*-grams, e.g. the imposition of first language structures, poor perception, or general tendencies toward simplification.

Third, one has to formulate hypotheses in advance, in order to test them. We may also use the method to 'data mine' for possible differences, but hypotheses to be tested must be formulated in advance (also see Section 6).

In spite of all these cautions, the method is quite powerful, and has been shown to be useful for at least data mining in our research on the English of Finnish immigrants in Australia. More on this in Section 4.

3 Rationale

Now we will discuss the reasons for, and assumptions behind our method, and after that, some of its features and options.

3.1 Why this method?

An important feature of our method is that it can identify not only deviant syntactic uses (errors), but also the overuse and underuse of linguistic structures, whose importance is emphasized by researchers on second-language acquisition (Coseriu, 1970; Ellis, 1994; de Bot *et al.*, 2005). According to these experts, it is misleading to consider only errors, as second-language learners likewise tend to overuse certain possibilities and tend to avoid (and therefore under-use) others. For example, de Bot *et al.* (2005) and Thomason (2001, p. 148) suggest that non-transparent constructions are systematically avoided even by very good second-language learners. And we expected to find this kind of SLA behaviour in our migrants' data.

This brings us to a second feature of the method: it can be applied to rather rough data, such as transcripts of speech or the writing of second-language learners. And as noted in the introduction it can be applied to any language, using any tag set for which there is a tagger or a tagged corpus available (*TnT* can be trained on any tagged corpus). We note that natural language processing work on tagging has found that larger tag sets result in lower accuracy (Manning and Schütze, 1999, p. 372 ff.). But since we aimed to contribute to the study of language contact and second-language learning, we chose a linguistically sensitive set designed by linguists, not computer scientists.

A third benefit of this method is that it delivers numerical data, and also provides significance levels with them. This allows one to go beyond the anecdotal and give a more rigorous grounding to dialectological or linguistic hypotheses. It could be important not only in the study of language contact, but also in the study of second-language acquisition. And it may be of more general linguistic interest, since contact effects are well-recognized confounders in the task of historical reconstruction. Numerical measures of syntactic difference may enable these fields to look afresh at many issues.

However, it is important to know that our method hinges on two assumptions, even if we think that they are reasonable and sound. We will introduce and discuss them now.

3.2 POS-*n*-grams show Syntax

The pivotal assumption behind our method is that POS-*n*-grams offer a good aggregate representation of syntax. And a first objection to it could be that POS-*n*-grams do not reflect syntax completely and that we thus should focus on full parse trees instead. However, since it is unlikely that researchers will take the time to hand-annotate large amounts of data, meaning we shall need automatically annotated data, we encounter a problem; that our parsers, the automatic data annotators capable of full annotation, are not yet robust enough for this task, especially for rougher data, such as spoken language (even the best score only about 90% per constituent on edited newspaper prose).

A second objection to the use of POS-*n*-grams could be that syntax concerns more than POS-*n*-grams. In response, we wish to deny that this is a genuine problem for the development of a measure of difference. We note that our situation is similar to other situations in which effective measures have been established. For example, even though researchers in first-language acquisition are very aware that syntactic development is reflected in the number of categories, and rules and/or constructions used, the degree to which principles of agreement are respected, etc., they are still in large agreement that the simple mean length of utterance (MLU) is an excellent measure of syntactic maturity (Ritchie and Bhatia, 1998). We therefore continue postulating that the measure we propose will correlate with syntactic differences as a whole, even if it does not measure them directly.

In fact, we can be rather optimistic about using POS-*n*-grams given the consensus in syntactic theory that a great deal of hierarchical structure is predictable given the knowledge of lexical categories, in particular given the lexical *head*. Sells (1982, Sections 2.2, 5.3, 4.1) demonstrates that this was common to theories in the 1980s (Government and Binding theory, Generalized Phrase Structure Grammar, and Lexical Function Grammar), and the situation has changed little in the successor theories (Minimalism and Head-Driven Phrase Structure Grammar). Even though the consensus of twenty years ago has been relaxed in recognition of the autonomy of constructions (Kay and Fillmore, 1999), syntactic heads still have a privileged status in determining a projection of syntactic structure. So it is likely that even individual POS-*n*-grams typical for a group of language users can give at least a good indication of what the differences in the underlying syntax might be.

Table 1 Performance of *TnT* on our data

N-grams	Full tagset (%)	Reduced tagset (%)
1-grams	81.2	86.7
2-grams	67.5	76.2
3-grams	56.1	66.6
4-grams	46.7	58.8
5-grams	39.0	51.8

Results per *n*-gram size for both the full and the reduced tagsets. Performance is better for smaller *n*-grams.

Table 2 Performance of *TnT* for the groups

N-grams	Youth (%)	Adults (%)
1-grams	81.7	80.9
2-grams	67.3	67.7
3-grams	56.0	56.2
4-grams	47.6	46.2
5-grams	40.4	38.2

Results of both groups, for the full tagset. There is consistent performance between groups.

3.3 Part of speech taggers are accurate enough

The method's second assumption is that POS-taggers are accurate enough for the POS-*n*-grams to reflect syntax at an aggregate level. First of all the facts: as noted, we used Thorsten Brants' *Trigrams 'n Tags* tagger. We used it with the tagset of the *TOSCA-ICE* consisting of 270 tags (Garside *et al*., 1997), of which 75 were never instantiated in our material. We trained the tagger on the British *ICE* corpus, which totals 1,000,000 words. Since our material was spoken English (see Section 4), we also did some experiments with training *TnT* on only the spoken half of the *ICE* corpus, but performance was better when using the whole corpus. Even then, using the British *ICE* corpus was suboptimal, as the material we wished to analyse was the English of Finnish emigrants to Australia, but we were unable to acquire sufficient tagged Australian material.

We tested the *TnT* tagger with a sample of 1,000 words from our material which was tagged by hand. We found that the tagger was correct for 81.2% of POS-tags. The accuracy is poor compared to newspaper texts, but we are dealing with conversation, including the syntactically imperfect conversation of non-natives here. Still, *n*-grams consist of multiple POS-tags, so performance is worse for larger *n*-grams; namely 56.1% for 3-grams, and even 39.0% for 5-grams. In addition, we also performed our small test using the reduced ICE tagset. It consisted of only 20 tags. For this performance was a bit better, namely: 86.7% for 1-grams, 66.6% for 3-grams, and 51.8% for 5-grams. See Table 1 for the full set of results.

So the performance of *TnT* is not that good for 3-grams and longer *n*-grams, meaning that our method is handicapped by half of the 3-grams of the full tagset being erroneous. But fortunately, we can fall back on a property of the statistics of large numbers here, namely that as the errors produced by automatic taggers are more or less random, they will cancel each other out, effectively annulling the influence of tagging errors. This cancelling of errors happens because in most cases we can expect the tagger to make similar mistakes in both groups, so the tagger favours neither of the groups. A tentative analysis confirmed this as we found that the tagging errors in the two groups seem to be of a similar kind, and also the error rates in both groups are very similar, as can be seen in Table 2 (also see the results in Section 4.3 for a further confirmation).

Thus, the results of the method in the form of a *p*-value and the POS-*n*-grams responsible for the difference, will be due mostly to the correctly tagged POS-*n*-grams. In addition, the use of the technique for data mining is unaffected by worries of tagging bias. Since in that case individually detected POS-*n*-grams are analysed and looked up in the corpus, so that any parsing errors that would show up in the results may then be corrected by hand or left out of the analysis.

4 Tangible Results

In this Section, we will summarize the research into the English of Finnish immigrants in Australia (Lauttamus *et al*., 2007). We will especially examine the trigrams responsible for the difference. We mainly test their usefulness for linguistic analysis here, in order to evaluate the method.

4.1 Application

We applied the procedure described in Section 2 to the *Finnish Australian English Corpus*, a corpus of interviews compiled in 1994 by Greg Watson of the University of Joensuu, Finland (Watson, 1995, 1996). The informants were all Finnish emigrants to Australia and they are classified into two groups in this report: (1) the adults (group '*a*', adult immigrants), who were 18 years or older upon arrival in Australia; (2) the juveniles (group '*j*', juvenile immigrant children of these adults), who were born in Finland and were all under the age of 18 years when they disembarked.

The goal of the research was to detect the linguistic sources of the syntactic variation between the two groups, and we examined the degree of what we call syntactic 'contamination' in the English of the adult speakers. Lauttamus *et al.* interpreted the findings from (at least) two perspectives, universal tendencies versus contact influence. The notion of 'universal' is concerned, not with hypotheses about Chomskyan universals, but rather with more general properties of the language faculty and natural tendencies in the grammar, called 'vernacular primitives' by Chambers (2003), pp. 265–66). To explain language usage by the groups, we also draw upon the strategies that second-language learners usually evince regardless of their mother tongue (Faerch and Kasper, 1983; Larsen-Freeman and Long, 1991; Ellis, 1994; Thomason, 2001)

4.2 Significance and trigrams

After applying the method, we found that the groups clearly differed in the distribution of POS trigrams they contain ($p < 0.0001$). This means that the difference between the original two sub-corpora was in the largest 0.01% of the Monte Carlo permutations, and thus highly significant. We also find genuinely deviant syntax patterns if we inspect the individual trigrams responsible for the difference between the two sub-corpora.

As noted in Section 2.5, only differences between the groups can be found using our method, and so we have no way to seek evidence of potential contamination of the juveniles' L2 acquisition relative to native speakers. Nevertheless, we can still deduce much from the POS trigrams that contributed to the differences (individual α-levels of 0.05). We limited our analysis to a practically random selection of 137 trigrams from the 308 trigrams found to be significant and typical for the adult speakers.[3] They were sorted by their relative typicality (see Section 5.3.2 for the normalization) and each interpreted using 20 random sentences from the corpus which contained them.

Table 3 Trigram counts for the eight findings

Index	Usage	Count
0	[useless]	24
1	Disfluent	90
2	Articles	39
3	No-Be	1
4	No-There	3
5	Pre-Negator	3
6	Relative-What	3
7	Present	3
8	Formulae	5
9	Between	3

The index number corresponds with the numbered findings in the text. Zero is used for trigrams that could not be interpreted (they have the '[useless]' label). Count is the number of occurrences among those looked at.

The findings concerning the adults can be described as follows. See Lauttamus *et al.* for an extended discussion.

(1) Overuse of hesitation phenomena (pauses, filled pauses, repeats, false starts, etc.) and parataxis (particularly with 'and' and 'but').

(2) Overuse (and underuse) of the indefinite and definite articles.

(3) Omission of primary (copula) 'be' and of the primary verb 'be' in the progressive (present and past).

(4) Underuse of the existential (expletive) 'there' and the anaphoric 'it' in subject position.

(5) Utterances where the negator 'not' is placed in pre-verbal position.

(6) Misuse of the pronoun 'what' as a relative pronoun or complementiser.

(7) Extension of the simple present (as opposed to the past tense and the progressive) to describe not only present but also past or future events.

(8) Indications of incorrect use of formulae such as 'that's' and 'what's' as in 'that's is a'.

(9) A tendency to place material between verbs and direct objects 'I don't like really any old age'.[4]

Table 3 shows a listing of the numbers of trigrams supporting each of the above claims.[5]

It should be noted that we have most evidence for disfluency and the over- and under-use of articles, though the other findings are also well supported. In addition, because this method also detects over- and under-use, the notion of avoidance does not therefore imply total absence of a feature in either group. See our earlier article (Lauttamus *et al.*, 2007) for a qualitative analysis.

4.3 Quantitative analysis

For a quantitative analysis, the *n*-grams can be sorted in two ways, namely by absolute and relative typicality. The first normalization (between-subjects normalization, Section 5.3.1) produces absolute data in the sense that the total number of occurrences of trigrams is decisive so that more frequent trigrams end up at the top. The second normalization (between-types normalization, Section 5.3.2) normalizes for frequency, and thus produces relative data, moving the relatively more typical trigrams to the top.

We start with a scatterplot (Fig. 1) of the distributions as produced by sorting the typical and significant trigrams by the two normalizations. Please note again that only the significant trigrams are sorted, and that exactly the same trigrams are sorted (but differently) on both lists.

As can be seen in the scatterplot, the more frequent trigrams are generally less typical and vice versa. Thus, the two ways of normalizing really do produce differently ordered lists of trigrams. To some extent, this relation is unsurprising as it is less likely that something very frequent is confined to only one group. Constructs that are frequent in a language are less likely to be relatively overused. Infrequent constructions, on the other hand, provide learners with too little evidence of their appropriate use. It is easier for them to rise to suspicious levels of relative frequency.

Our second graph (Fig. 2) shows the cumulative percentage of interpretable trigrams found at each rank position. This graph is much like recall graphs as used in Information Retrieval. For example, in the 'absolute' data, approximately 40% of the interpretable trigrams are found in the top 120.

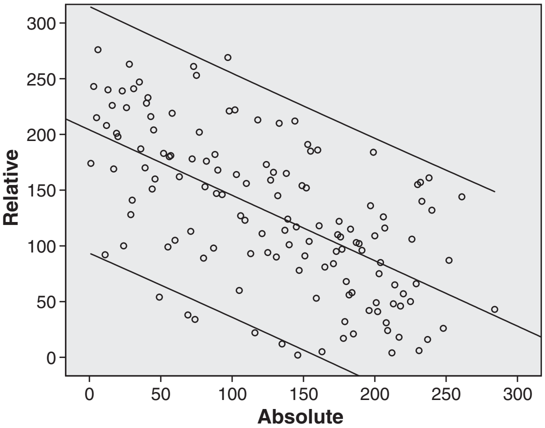

Fig. 1 Distribution of the 137 randomly selected trigrams ranked according to absolute values on the *x*-axis (only between-subjects normalized) and according to relative values on the *y*-axis (normalized for frequency, between-types normalized). The parallel lines denote 1 standard deviation. More frequent trigrams are less typical and vice versa.

What it shows is slightly surprising, namely that the absolute data, that is the data normalized with just the between-subjects normalization (the black line) is better than the relative (normalized for frequency, the grey line) data. The difference is modest for the whole collection of useful trigrams, but when discarding the first two categories (both disfluency and articles), the results are more alike. This suggests two things, first that the overuse of articles occurs in frequent, and not very typical trigrams, and second, that the same might be true for hesitations, if perhaps to a lesser extent. While not conclusive, the tendency suggests it might be a good idea to look at the absolute data first, especially if many *n*-grams are found or when time constraints are strict.

Lastly, we also hand-checked the performance of the *TnT* tagger for the 20 examples found with each of the analysed trigrams from the top 308 list. While doing this, we examined a window of three words on each side of the trigram, so the context was taken

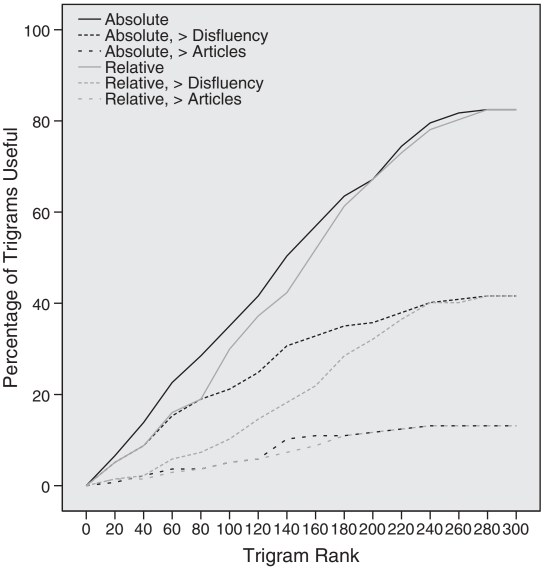

Fig. 2 Recall of interpretable trigrams as a function of both the absolute (only between-subjects normalized) and relative (normalized for frequency, between-types normalized) sort order. Absolute data works better than relative data. Most of this difference is in overuse of articles and disfluency, which occurs in frequent, and not very typical (not typical for a group) trigrams.

into account. The results of this are very promising, because when compared to the performance of *TnT* on the whole corpus, performance for the significant *n*-grams is up by almost 20 percentage points; 76.3% as opposed to the overall 56.5%.

This means that the errors as introduced by the tagger are to a great extent indeed random noise that does not disturb our method, confirming what we argued for in Section 3.3. Note also that 76.3% approaches the ceiling of *TnT*s 81.2% performance on 1-grams very closely. This means both that the method will likely work even better on cleaner data, and that even though we might still be missing something due to systematic tagging errors, it is unlikely to be very much, or to cause false overall significance.

Thus, the evaluation of our method via our experiment on the corpus containing the English of Finnish emigrants to Australia, is promising in that the method works well, both in distinguishing two different groups of speakers, and in highlighting relevant syntactic deviations between the two groups.

5 Statistical Theory

The test we need has to be able to check whether the differences between frequency vectors containing 8,300 elements are statistically significant, and how significant the individual differences are. Also, it has to be a method that will not find significance for tiny differences due to the large number of variables alone. Fortunately, permutation tests fulfil these requirements when used wisely (Good, 1995; Moore and McCabe, 2005). Kessler (2001) contains an informal introduction for an application within linguistics.

5.1 Permutation tests

As noted, the fundamental idea behind permutation tests with Monte Carlo permutation is very simple, but the mathematics is very simple as well: we measure the difference between the two original sets in some convenient fashion, obtaining $\delta(A, B)$. We then apply the Monte Carlo permutations (random permutations), which means extracting two sets of the same size as the base sets at random from $A \cup B$. We call these two random sets A_1, B_1, and we calculate the difference between these two in the same fashion, $\delta(A_1, B_1)$, recording if $\delta(A_1, B_1) \geq \delta(A, B)$, i.e. if the difference between the two randomly selected subsets from the entire set of observations is as extremely or even more extremely different than the original sets.

If we repeat this process of permutations, say, 10,000 times, then counting the number of times we obtain differences at least as extreme as in the base case, allows us to calculate how strongly the original two sets differ from a chance division with respect to δ. In that case, we may conclude that if the two sets were not genuinely different, the original division into A and B was likely to the degree of $p = n/10{,}000$. Put into more standard hypothesis testing terms, p is the p-value, i.e. the probability of seeing the difference between the two sets if there is no difference in the populations they represent. Based on this value, we may reject (or retain) the null hypothesis that there is no significant difference between the two sets.

Permutation tests are quite different from parametric tests such as the (M)ANOVA and the *t*-test in that they make fewer assumptions and are applicable to more kinds of data. They do not require the data to be normally distributed (to conform to any particular probability distribution), or to be homoscedastic (have regular and finite variance). And this is good news as most linguistic research violates at least one of these requirements.

Permutation tests can also be more sensitive than parametric tests such as *t*-tests, as the latters estimations of significance depend on normality assumptions, etc., while permutation tests are tailored to the data. The permutation test implicitly generates the equivalent of a distribution table during permutation. Kemperthorne and Fisher remark: ‘conclusions [of the ANOVA] have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method [the permutation test]’ (Edington, 1987, p. 11). Permutation tests are also easier to understand and more transparent than many parametric tests, at least for most non-mathematicians.

5.2 Exchangability and relevance

Permutation tests have only two real requirements which are relatively straightforward. First, one needs to be alert about exactly what relationships between the datasets are being measured by ones test and its permutations. What is important here is the exchangeability of elements under the null-hypothesis, or in other words that there is no interdependence between the elements that are being permutated. We can best illustrate this by explaining our decision to measure the difference in syntax by permutating the speakers in the two groups, and not the *n*-grams or the sentences—as we did previously (Nerbonne and Wiersma, 2006). Because if we permute separate *n*-grams then we might—besides the differences between the groups—also measure the syntactic relationships between (overlapping) *n*-grams within sentences. By permutating sentences of multiple different speakers across two groups we might also measure differences between the personal syntactic styles of the subjects, instead of just those differences that are caused by them being members of the two groups. Only the speakers can be considered to be truly independent and exchangable.

Second, as always, there is an important difference between statistical significance and effect size. A statistically significant result does not necessarily indicate a big difference. Significance measures the likelihood of a difference being due to chance. Very small, but consistent differences can always be made significant by increasing the size of the dataset. For example, if height were related to presentation skills, but only if averaged over millions of people, then knowing that a a presenter is tall, would not tell you very much about the quality of the upcoming presentation. The problem in this example is that the relationship is not strong enough. Sheer corpus size will lead to higher estimates of statistical significance when using permutation tests, and having a bigger corpus will always make it easier to find significant differences (Edington, 1987; Agresti, 1996).[6]

In short: when used with the necessary care, permutation tests are a very suitable tool for finding significant syntactical differences and the POS-*n*-grams that contribute to this difference.

5.3 Normalizations

Each measurement of difference that is part of the permutation test—whether the difference is between the original two samples or between two samples which arise through permutations—is applied to the collection of POS-*n*-gram frequencies once these have been normalized.

We first describe the normalization that is required. This normalization is called the between subjects normalization. It corrects for a whole range of factors that could otherwise violate the requirements of permutation tests, and it is thus quite conservative and robust. It normalizes for differences in the size of texts, instead of for variations in sentence length, like the in-permutation normalization described in our previous article on the method (Nerbonne and Wiersma, 2006), which was both more complicated and a bit less conservative.

The second normalization we used is called the between-types normalization, and it normalises for differences in frequency between *n*-gram types. Using it is optional and its purpose is mainly the elimination of frequency as a factor, allowing one to detect significance if the typical *n*-grams are the less frequent ones. It is also different from the normalization that we used for the same purpose before (the between-permutations normalization). The difference being that the between-types normalization is much simpler to calculate for the output of the between-subjects normalization, and that part of its function is now delegated to a third normalization.

The between-groups normalization is the third normalization, and it has a really different function from the others as its output is not used as input for a measure or a permutation test, but is solely meant for detecting which of the two groups a given *n*-gram is typical for. So it can only be used after individual *n*-grams were selected by using one of the other normalizations. It normalizes for group size, so it can detect what constitutes over- or underuse, even if the groups are of different sizes.

5.3.1 *The between-subjects normalization*

The between-subjects normalization is, as noted, applied to texts belonging to single authors or speakers. Its purpose is to make the collection of POS-*n*-grams that make up the texts of a speaker comparable to that of other speakers so that no one subject has more influence on the groups vector than any other. This is also important for permutating, as it ensures that the sizes of the groups in terms of *n*-grams counts will not change when authors are permutated between them, so the requirement of exchangeability (constant group sizes) is not violated. The normalization functions, in short, by correcting for the size of the text by the speaker as measured by the number of POS-*n*-grams inside it.

It has the possible disadvantage that it might not always be possible, easy or feasible to split up a corpus according to authorship. For example, some texts, such as those in the Bible, have multiple or contested authors. Such problems may be solved, we believe, but we omit discussions in the interest of brevity.

For the between-subjects normalization, we first collect from the tagger a sequence of counts c_i of POS-tag-*n*-grams (index *i*, total number *n*) for each subject (*c*). Giving us one vector ($\mathbf{c}^s$) per subject.

$$\mathbf{c}^s \quad = \quad < c_1^s, c_2^s, \ldots, c_n^s >$$

After that, we normalize for total number of POS-*n*-grams produced by a given subject. We do this by calculating the sum of the subject's *n*-gram counts, and then dividing each of his individual *n*-gram counts by it. This gives us an *n*-gram fraction vector ($\mathbf{f}^s$) for the subject.

$$N^s = \sum_{i=1}^{n} c_i^s$$

$$\mathbf{f}^s \quad = \quad < \ldots, f_i^s (= c_i^s / N^s), \ldots > \quad \sum_{i=1}^{n} f_i^s = 1$$

After we have done this for all the subjects, we have a list (**F**) that contains vectors, namely the fraction vectors of all subjects (with *m* as the total number of subjects).

$$\mathbf{F} = < \mathbf{f}_1^s, \mathrm{f}_2^s, \ldots, \mathrm{f}_m^s >$$
$$m = |j| + |a|$$

Then we use the fraction vectors in this list as the elements we permutate, instead of *n*-grams or sentences, effectively permutating speakers. We permutate these subjects between two groups, which we shall refer to as juveniles ('*j*') and adults ('*a*') as we did in our example research. Each permutation results in two lists—one for each group—containing the fraction vectors of the subjects ending up in that group ($\mathbf{f}^{sj}$ for a subject vector in the juveniles group). So for each permutation, including the base case, we produce two lists of fraction vectors ($\mathbf{F}^\mathbf{j}$ and $\mathbf{F}^\mathbf{a}$).

$$\begin{aligned} \mathbf{F}^\mathrm{j} &= < \mathrm{f}_1^{sj}, \mathrm{f}_2^{sj}, \ldots, \mathrm{f}_{|j|}^{sj} > \\ \mathbf{F}^\mathrm{a} &= < \mathrm{f}_1^{sa}, \mathrm{f}_2^{sa}, \ldots, \mathrm{f}_{|a|}^{sa} > \end{aligned}$$

As the last step of this normalization for each group, we sum the fraction vectors of all subjects that ended up in that group (summing the fractions per *n*-gram for different authors in the group, not across different *n*-grams), giving us a vector holding per-*n*-gram sums for each group: s^j and s^a (note this last step is also done for for all permutations).

$$\begin{aligned} \mathrm{s}^\mathrm{j} &= \sum_{\mathrm{i}=1}^{|\mathrm{j}|} \mathbf{F}_\mathrm{i}^\mathrm{j} \\ \mathrm{s}^\mathrm{a} &= \sum_{\mathrm{i}=1}^{|\mathrm{a}|} \mathbf{F}_\mathrm{i}^\mathrm{a} \end{aligned}$$

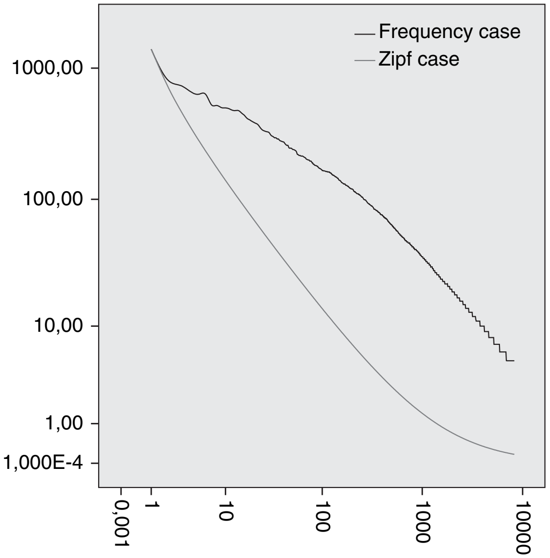

Fig. 3 Distribution of the trigrams in the FAEC corpus in a log graph, with the ideal Zipf distribution in grey. Trigrams largely follow Zipf's law, so a small number of frequent trigrams can make a large difference if the data is not normalized for this.

The pair of vectors it produces at each permutation (s^j and s^a) can, besides being already directly usable when no further normalization is done for frequency, also be used as input for the second normalization, the between-types normalization.

5.3.2 *The between-types normalization*

The purpose of the between-types normalization is the removal of the influence of absolute frequencies in *n*-gram counts. This is useful because as can be seen in the graph for trigrams (Fig. 3), the frequency of POS-*n*-gram types follows Zipf's law, and thus a few very frequent (perhaps uninteresting) *n*-grams could have too much influence on the reported significance of the difference between the two groups.

So this normalization allows one to find significance, regardless of frequency if there are enough *n*-gram types that are typical for the two original sub-corpora,

It is similar to the second step of the subjects normalization (the fractions per author), except that it is based on the total count for each *n*-gram type: for each POS-*n*-gram type *i* in each group (sub-corpus) $g \in \{a, j\}$, the summed count of the group s_i^g is divided by the total count (both as provided by the subjects normalization, so they are summed fractions, not raw counts) for that type s_i. The outer part of the formula for $\mathbf{t}^g$ is (where s_i^j is the count of the POS-*n*-gram *i* which occurs in group *j*):

$$\begin{aligned}\mathbf{t}^j &= < \ldots, t_i^j (= s_i^j / s_i), \ldots > \\ \mathbf{t}^a &= < \ldots, t_i^a (= s_i^a / s_i), \ldots >\end{aligned}$$

The inner part of the formula, the total count for each *n*-gram (s_i), can be calculated for the whole vector ($\mathbf{s}$) as follows:

$$\mathbf{s} = < \ldots, s_i (= s_i^j + s_i^a), \ldots >$$

N-grams with large summed authors' fraction counts are those with high frequencies in the original sub-corpora. The normalization under discussion strips away the role of frequency and thus allows us to find significance if there are typical *n*-grams, even if they are less frequent.

The values thus normalized will be 0.5 on average when the groups are of the same size. They will be skewed in the direction of the larger group when one of the groups is bigger than the other. Therefore, the output of this normalization cannot directly be used to determine if there is under- or over-use (as 0.3 can be over-use for a small group, and extreme under-use for a bigger one).

It is primarily meant for determining the overall (corpus-wide) *p*-value without regard to frequency. It has no influence on the *p*-values reported for individual *n*-grams, because it is only a linear transformation when looked at per *n*-gram. In addition, we use it for sorting the *n*-grams that were found to be significant, as it will cause infrequent, but typical *n*-grams to move to the top (see Section 2.4).

The normalization is applied to the base case and all subsequent permutations, and the pair of arrays it produces ($\mathbf{t}^j$ and $\mathbf{t}^a$) is ready to be used as input for a measure.

5.3.3 *The between-groups normalization*

The between-groups normalization is applied to the output of the first normalization (the between-subjects normalization) and it is meant for detecting whether an *n*-gram is being overused in one group or the other. The average normalized value will be l, with cases of overuse in the group having a normalized value of bigger than l, and cases of underuse having a value smaller than l. It is calculated by dividing the per-*n*-gram count by the average value for that *n*-gram in that group. It works because the average value for each of these *n*-grams across permutations is equal to the expected value for a group of that size under the null-hypothesis.

More formally, it is arrived at in the following way: for each POS-*n*-gram i in each group (sub-corpus) $g \in \{a, j\}$, the summed count s_i^g is divided by the average of the summed count (both as provided by the between-subjects normalization) for that type in that group (across all permutations): $\overline{s_i^g}$. The outer part of the formula for $\mathbf{o}^g$ is:

$$\mathbf{o}^{\mathrm{j}} = < \ldots, o_i^j(= s_i^j/\overline{s_i^j}), \ldots >$$
$$\mathbf{o}^{\mathrm{a}} = < \ldots, o_i^a(= s_i^a/\overline{s_i^a}), \ldots >$$

Now for large numbers of permutations, the inner part of the formula, namely the average count ($\overline{s_i^g}$), can be calculated as $(s_i^j + s_i^a)/2$ when the groups are of equal sizes, and as follows when the groups differ in size (where N^j and N^a are the number of subjects in the juveniles, and adults group, respectively):

$$\overline{\mathbf{s}^{\mathrm{j}}} \;=\; <, \overline{s_i^j}(= (s_i^j + s_i^a) \cdot N^j/(N^j + N^a)), >$$
$$\overline{\mathbf{s}^{\mathrm{a}}} \;=\; <, \overline{s_i^a}(= (s_i^j + s_i^a) \cdot N^a/(N^j + N^a)), >$$

The values thus normalized will be l on average across permutations, and thus when summed for large corpora, be equal to the number of POS-*n*-gram types.

$$\sum_{i=1}^{n} o_i^j = n$$

$$\sum_{i=1}^{n} o_i^a = n$$

We firmly note again that the per-POS-*n*-gram values output by this normalization are only useful for detecting over- and under-use when used together with information on per-POS-*n*-gram statistical significance as provided by one of the other normalizations.

First of all, it cannot be used to find typical or extreme POS-*n*-grams without consideration for significance, because infrequent *n*-grams are especially likely to have high values with respect to $\overline{s_i^g}$. For example, an *n*-gram occurring only once, in one sub-corpus, will get a value of $1/0.5 = 2$ (as it is indeed very typical for this sub-corpus), while with a count of one it clearly will not be statistically significant (moving between equally sized sub-corpora with a chance of 50% during permutations).

Second, this normalization is not suitable to generate per *n*-gram *p*-values itself when group sizes differ, because then the output will not be symmetrical. To illustrate this, we can look at our single *n*-gram again: as an *n*-gram occurring in a *smaller* sub-corpus can produce very big normalized values, such as $1/0.2 = 5$, and one in a bigger sub-corpus will always produce smaller values, such as $1/0.8 = 1.25$. This can lead to false significance by introducing fluctuations in the normalized group sizes, and by turning two-sided measures, such as *RSquare*, into one-sided ones (more on *RSquare* in Section 5.4).

When these two things are kept in mind, the between-groups normalization can be used for assigning the detected POS-*n*-grams to the list of the group which overuses them. Just as the previous normalizations, this normalization is also applied to the base case and all subsequent permutations, and for each it produces a pair of arrays ($\mathbf{o}^{\mathrm{j}}$ and $\mathbf{o}^{\mathrm{a}}$).

5.4 Measures

The choice of vector difference measure, e.g. cosine versus χ^2, does not affect the proposed technique greatly, and alternative measures can easily be used. Accordingly, we have worked with both cosine and two measures inspired by the recurrence (R) metric introduced by Kessler (2001, p. 157 and further). We also call these measures R and *RSquare*. The advantage of the R and *RSquare* metrics is that they are transparent since they are simple aggregates that allow one to easily see how much each *n*-gram contributed to the difference. We also used these measures to calculate a separate *p*-value per *n*-gram.

Our R is calculated as the sum of the differences of each cell with respect to the average for that cell. If we have collected our data into two vectors ($\mathbf{s}^{\mathrm{j}}$,

$\mathbf{s}^a$), and if *i* is the index of a POS-*n*-gram, *R* for each of these two groups is equal, and it simply looks as follows.

$$R = \sum_{i=1}^{n} \left| s_i^j - \overline{s_i^g} \right|$$

With the average between the two groups for each *n*-gram cell ($\overline{s_i^g}$) being.

$$\overline{s_i^g} = (s_i^j + s_i^a)/2$$

The *RSquare* measure emphasizes a few large differences more than many small ones, and it is simply the square of *R*, and thus calculated in this way (with $\overline{s_i^g}$ being the same as for *R*).

$$RSquare = \sum_{i=1}^{n} (s_i^j - \overline{s_i^g})^2$$

Both *R* and *RSquare* are two-sided measures. This means that the same *R(Square)* value is obtained when an *n*-gram has more than the average value to a certain extent (say *x*), as when it has less than the average to this same extent (also *x*). So tests using them are two-tailed (conservative). Both very typical and atypical *n*-grams are detected as significant. Thus, assignment to the groups for which they are typical is done as a separate step using the between-groups normalization (see Section 5.3.3).

All in all, we had best results with the *RSquare* measure and thus have used it for our example research.

6 Further Research

First, we will go into possible applications of the method, and then we will give some suggestions for testing, analysis, and improvement of the method.

6.1 Applications of the method

As already shortly mentioned in the introduction, the method as presented here can be applied to many datasets. It can offer quantitative answers to many questions other than our questions on the two generations of immigrants. More specific questions could be asked such as the time course needed for second-language acquisition, the relative importance of factors influencing the degree of difference such as the mother tongue of the speakers, other languages they know, the length, and time of their experience in the second language, the role of formal instruction, etc. And beyond this, comparisons could be made between dialects and variants of languages (such as the various Englishes as collected in the ICE corpora). Also different discourses can be compared, such as the language of lawyers or academics as compared to that of laymen or students.

Allthough we expect our technique to be especially useful to corpora involving speech or disfluent language, it might also be applied to the analysis of literary styles such as the syntactic difference between romance and detective novels. The change of syntax through time could also be followed, as reflected in novels or newspapers published in different decades or even different centuries. In addition, it might be used to track traces of the grammar of a source language (say Ancient Greek) in modern translations of classical texts. As long as a common tagset exists or can be devised for a corpus of comparable material, and the comparison answers one or more meaningful questions, the method can be used.

It might even be used in teaching, so second-language learners, aspiring creative writers, or journalism students can be shown what grammatical structures they—either collectively or individually—are over- or under-using compared to native speakers, famous authors, or acclaimed journalists. If the method is confirmed to work for small corpus sizes and/or is improved and calibrated sufficiently, many practical applications may be realizable.

6.2 Analysis and improvement of the method

The method should be analysed in more depth with respect to the corpus size at which it is effective in finding genuine differences. A minimum size should be established (smallest corpus size at which a significant difference may be found if there is one). It would also be valuable to analyse the various types of errors in analysis that the method might produce. So far, we found better tagger performance for the significant *n*-grams, as noted in Sections 3.3 and 4.3. Another opportunity for analysis might be the testing of the method with various tag sets and *n*-gram sizes: especially tag sets of various precisions, and for larger *n*-grams.[7]

The method used here exploratory can be used straightforwardly in hypothesis testing, e.g. by identifying a class of POS-*n*-grams for which one predicts to find significant differences. Lauttamus, Nerbonne, and Wiersma (to appear) examine the hypotheses that (I) filled pauses infect a great deal of L2 speech, but (II) that they do not exhaust the significant differences. One might also employ a cross-validation design in which the corpus is repeatedly split into parts, say 20% stands to 80%, where the second part is used to mine for typical *n*-grams which may then be confirmed by applying the method to the first part. Another improvement would be to add a stronger protection against family-wise errors, such as partial null hypothesis family-wise errors (part of null-hypothesis being true, as for separate *n*-grams). In the literature on neuro-imaging, where comparisons between many data points also have to be made, proposals and solutions have recently been put forth (Nichols and Holmes, 2002; Nichols and Hayasaka, 2003)

In addition, it could be useful to experiment with various measures, and especially to calibrate it so we would have a standardized, overall difference-value in addition to a *p*-value. When parsers become more accurate, one could go beyond POS-tags, especially for cases where tagging accuracy might be less an issue, such as in newspaper text or novels as Sanders and others have shown (Baayen, 1996; Hirst and Feiguina, 2007; Sanders, 2007).

7 Conclusion

Weinreich (1968) regretted that there was no way to 'measure or characterize the total impact one language on another in the speech of bilinguals' (1968, p. 63), and speculated that there could not be. This article has proposed a way of going beyond counts of individual phenomena to a measure of aggregate syntactic difference.

The technique proposed follows Aarts and Granger (1998) in using part-of-speech *n*-grams. We argue that such lexical categories are likely to reflect a great deal of syntactic structure given the tenets of linguistic theory according to which more abstract structure is, in general, projected from lexical categories. We went beyond Aarts and Granger in showing how entire vectors of POS *n*-grams may be used to characterize aggregate syntactic distance, and in particular by showing how these, and individual *n*-gram counts can be analysed statistically.

The technique was implemented using an automatic POS-tagger, several normalizations and permutation statistics, and it was shown to be effective on the English of Finnish immigrants to Australia. We were able to detect various forms of interference of their L1 (Finnish) on the English of the adult speakers, such as overuse of articles and placing not in pre-verbial position, as well some due to universal contact influences, such as the overuse of hesitation phenomenon.

The *ComLinToo*, the software implementing the method, including the normalizations, is freely available. It is developed to allow easy application to other datasets, and generalization to *n*-grams of any size.

There are many possibilities for future research. First of all, it can be applied to a wide range of datasets and used to answer many questions, such as factors influencing language learning, comparisons between discourses and literary styles, and maybe even teaching. Second, the method could be further analysed, testing it with various corpus, tag set and *n*-gram sizes, and doing more qualitative analysis. Lastly, there are also many opportunities to improve the method, by adding and evaluating more statistical safe-guards, by experimenting with various measures, and by using *chunking* or *parsing*, instead of POS-tags, especially for cleaner data.

Thus, while we fall short of Weinreich's goal of assaying 'total impact' in that we focus on syntax, we do take a large step in this direction by showing how to aggregate, test for statistical significance, and examine the syntactic structures responsible for the difference.

Acknowledgements

We are grateful to Lisa Lena Opas-Hänninen and Pekka Hirvonen of the University of Oulu, who made the data available and consulted extensively on its analysis. We also thank audiences at the 2005 *TaBu Dag*, Groningen; at the Workshop *Finno-Ugric Languages in Contact with English II* held in conjunction with *Methods in Dialectology XII* at the *Université de Moncton*, August 2005; the Sonderforschungsbereich 441, 'Linguistic Data Structures', Tübingen in January 2006; the Seminar on Methodology and Statistics in Linguistic Research, University of Groningen, Spring, 2006; the *Linguistic Distances Workshop* at the *Coling/ACL 2006* in Sydney, July 2006; and the *Thirteenth International Conference on Methods in Dialectology*, Leeds, August 2008, and especially Livi Ruffle, for useful comments on our earlier article.

References

Aarts, J. and Granger, S. (1998). Tag Sequences in Learner Corpora: a Key to Interlanguage Grammar and Discourse. In Granger, S. (ed.), *Learner English on Computer*. London: Longman, pp. 132–41.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley.

Baayen, H., van Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11: 121–32.

Bot, de K., Lowie, W., and Verspoor, M. (2005). *Second Language Acquisition: An Advanced Resource Book*. London: Routledge.

Brants, T. (2000). TnT - a statistical part of speech tagger. In *6th Applied Natural Language Processing Conference*. Seattle: ACL, pp. 224–31.

Chambers, J. (2003). *Sociolinguistic Theory: Linguistic Variation and its Social Implications*. Oxford: Blackwell.

Coseriu, E. (1970). *Probleme der kontrastiven Grammatik*. Düsseldorf: Schwann.

Edington, E. S. (1987). *Randomization Tests*. New York: Marcel Dekker Inc.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Faerch, C. and Kasper, G. (1983). *Strategies in Interlanguage Communication*. London: Longman.

Garside, R., Leech, G., and McEmery, T. (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London/New York: Longman.

Good, P. (1995). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22: 405–17.

Kay, P. and Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: the what's x doing y construction. *Language*, 75: 1–33.

Kessler, B. (2001). *The Significance of Word Lists*. Stanford: CSLI Press.

Larsen-Freeman, D. and Long, M. H. (1991). *An Introduction to Second Language Acquisition Research*. London: Longman.

Lauttamus, T., Nerbonne, J., and Wiersma, W. (2007). Detecting syntactic contamination in emigrants: the English of Finnish Australians. *SKY Journal of Linguistics*, 20: 273–307.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Moore, D. S. and McCabe, G. P. (2005). *Introduction to the Practice of Statistics*. New York: W. H. Freeman.

Nelson, G., Wallis, S., and Aarts, B. (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nerbonne, J. and Wiersma, W. (2006). A Measure of Aggregate Syntactic Distance. In Nerbonne, J. and Hinrichs, E. (eds), *Linguistic Distances*. PA: ACL, Shroudsburg, pp. 82–90.

Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12: 419–46.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15: 1–25.

Odlin, T. (1989). *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

Odlin, T. (1990). Word-order Transfer, Metalinguistic Awareness and Constraints on Foreign Language

Learning. In VanPatten, B. and Lee, J. (eds), *Second Language Acquisition/Foreign Language Learning*. Clevedon, UK: Multilingual Matters, pp. 95–117.

Odlin, T. (2006). Could a contrastive analysis ever be complete? In Arabski, J. (ed.), *Cross-linguistic Influence in the Second Language Lexicon*. Clevedon, UK: Multilingual Matters, pp. 22–35.

Odlin, T. (2009). Methods and inferences in the study of substrate influence. In Filppula, M., Klemola, J., and Paulasto, H. (eds), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*. New York: Routledge, pp. 265–79.

Poplack, S and Sankoff, D. (1984). Borrowing: the synchrony of integration. *Linguistics*, 22: 99–136.

Poplack, S., Sankoff, D., and Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26: 47–104.

Ritchie, W. C. and Bhatia, T. K. (1998). *Handbook of Child Language Acquisition*. San Diego: Academic.

Sampson, G. (2000). A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics*, 5: 53–68.

Sanders, N. C. (2007). Measuring Syntactic Differences in British English. In *Proceedings of the Student Research Workshop*. Madison: Omnipress, pp. 1–7.

Sells, P. (1982). *Lectures on Contemporary Syntactic Theories*. Stanford: CSLI.

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

Thomason, S. G. (2001). *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press.

van Coetsem, F. (1988). Loan phonology and the two transfer types in language contact. *Publications in Language Sciences*. Dordrecht: Foris Publications, p. 27.

Watson, G. (1995). A Corpus of Finnish-Australian English: a Preliminary Report. In Muikku-Werner, P. and Julkunen, K. (eds), *Kielten Väliset Kontaktit*. Jyväskylä: University of Jyväskylä, pp. 227–46.

Watson, G. (1996). The Finnish-Australian English corpus. *ICAME Journal: Computers in English Linguistics*, 20: 41–70.

Weinreich, U. (1968). *Languages in Contact*. The Hague: Mouton.

Westfall, P. H. and Young, S. S. (1993). *Resampling Based Multiple Testing: Examples and Methods for p-value Adjustment*. New York: John Wiley & Sons, Inc.

Notes

[1] List of taggers: http://www-nlp.stanford.edu/links/statnlp.html#Taggers; CLAWS & TreeTagger are on this list; Tsujiis Tagger: http://www-tsujii.is.s.u-tokyo.ac.jp/∼tsuruoka/postagger/; ALPINO: http://www.let.rug.nl/ vannoord/alp/Alpino/ (only for Dutch).

[2] The *Computational Linguistics Toolset*, which was developed for the method, is available from http://en.manta.logilogi.org/Wybo_Wiersma/User/Com_Lin_Too=Wybo_Wiersma_22. This web page also contains an introduction to their usage and information on the other auxiliary tools.

[3] Originally, it was the top-176 obtained using our earlier, slightly less robust normalization.

[4] Note that this is an example where the technique detects a violation involving a constituent (direct object) and not merely low-level word categories.

[5] A full list of all trigrams with examples from the corpus can be downloaded at http://wybowiersma.net/pub/fiauimenre/separate-table.pdf. In this list, the 'usages' corresponds to the 'usage' below. The corpus is now available here: http://wybowiersma.net/pub/fiauimenre/faec.tgz.

[6] In our 2006 paper, we formulated this less clearly and conservatively. In fact, we thought then that corpus size could be ignored, which is wrong.

[7] There appears to be a bandwidth between 1- and 5-grams at which it is easier to find significance, but apart from noting decreased significance at the edges, we have not yet been able to map it more precisely.