A Measure of Aggregate Syntactic Distance

John Nerbonne and Wybo Wiersma
Alfa-informatica, University of Groningen
P.O.Box 716, NL 9700 AS Groningen, The Netherlands
j.nerbonne@rug.nl & wybo@logilogi.org

Abstract

We compare vectors containing counts of trigrams of part-of-speech (POS) tags in order to obtain an aggregate measure of syntax difference. Since lexical syntactic categories reflect more abstract syntax as well, we argue that this procedure reflects more than just the basic syntactic categories. We tag the material automatically and analyze the frequency vectors for POS trigrams using a permutation test. A test analysis of a 305,000 word corpus containing the English of Finnish emigrants to Australia is promising in that the procedure proposed works well in distinguishing two different groups (adult vs. child emigrants) and also in highlighting syntactic deviations between the two groups.

1 Introduction

Language contact is a common phenomenon which may even be growing due to the increased mobility of recent years. It is also linguistically significant, since contact effects are prominent in linguistic structure and well-recognized confounders in the task of historical reconstruction. Nonetheless we seem to have no way of assaying the aggregate affects of contacts, 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, 1953, p. 63)

This paper proposes a technique for measuring the aggregate degree of syntactic difference between two varieties. We shall thus attempt to measure the "total impact" in Weinreich's sense, albeit with respect to a single linguistic level, syntax.

If such a measure could be developed, it would be important not only in the study of language contact, but also in the study of second-language acquisition. A numerical measure of syntactic difference would enable these fields to look afresh at issues such as the time course of 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. It would make the data of such studies amenable to the more powerful statistical analysis reserved for numerical data.

Naturally we want more than a measure which simply assigns a numerical value to the difference between two syntactic varieties: we want to be able to examine the sources of the difference both in order to win confidence in the measure, but also to answer linguistic questions about the relative stability/volatility of syntactic structures.

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. People MAINTAINING their language tend to adopt new lexical items from a contact language, but this only has structural consequences as the lexical items accumulate. Thus we hear radically different English used in immigrant

communities in the English-speaking world, but the natives in contact with these groups do not tend to modify their language a great deal. This suggests that we should concentrate on those switching as we begin to develop measures of aggregate difference.

Poplack and Sankoff (1984) introduced techniques for studying lexical borrowing and its phonological effects, and and Poplack, Sankoff and Miller (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 suggest a means of measuring the aggregate difference between two varieties, and show how we can test whether that difference is statistically significant.

2 Syntactic Footprints

In this section we justify using frequency profiles of trigrams of part-of-speech (POS) categories as indicators of syntactic differences. We shall first automatically tag second-language speakers' corpora with syntactic categories:

Oh	that	's	a	just	a
INT	PRON	COP	ART	EXCL	ART
fun	in	a	,	Helsinki	
N-COM	PREP	ART	PAUSE	N-PROP	

We then collect these into overlapping triples (trigrams). The tag-trigrams include triples such as INT-PRON-COP and PRON-COP-ART.

We consider three possible objections to proceeding this way. First, one might object that unigrams, bigrams, also should be compared. We are in fact sympathetic to the criticism that n-grams for $n \neq 3$ should also be compared, at least with an eye toward refining the technique, and we have performed experiments with bigrams and with combinations of n-grams for larger n, but we restrict the discussion here to trigrams in order to simplify presentation. Second, our choice of part-of-speech categories may bias the results, since other research might use other POS categories, and third, that POS trigrams do not reflect syntax completely. We first develop these last two objections further, and then explain why it is reasonable to proceed this way.

Ideally we should like to have at our disposal the syntactic equivalent of an international phonetic alphabet (IPA, 1949), i.e. an accepted means of noting (an interesting level of) syntactic structure for which there was reasonable scientific consensus. But no such system exists. Moreover, the ideal system would necessarily reflect the hierarchical structure of dependency found in all contemporary theories of syntax, whether directly based on dependencies or indirectly reflected in constituent structure. Since it is unlikely that researchers will take the time to hand-annotate large amounts of data, meaning we shall need automatically annotated data, this leads to a second problem, viz., that our parsers, the automatic data annotators capable of full annotation, are not yet robust enough for this task. (Even the best score only about 90% per constituent on edited newspaper prose.)

We have no solution to the problem of the missing consensual annotation system, but we wish to press on, since it will be sufficient if we can provide a measure which correlates strongly with syntactic differences. We note that natural language processing work on tagging has compared different tag sets, noting primarily the obvious, that larger sets result in lower accuracy (Manning and Schütze, 1999, 372ff.). Since we aim here to contribute to the study of language contact and second-language learning, we shall choose a linguistically sensitive set, that is, a large set designed by linguists. We have not experimented with different tagsets.

With regard to the second objection, the fact that syntax concerns more than POS trigrams, we wish to deny that this is a genuine problem for the development of a measure of difference. We note that our situation in measuring syntactic differences 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 and government are respected, the fidelity to adult word order patterns, etc., still they are in large agreement that the very simple MEAN LENGTH OF UTTERANCE (MLU) is an excellent measure of syntactic maturity (Ritchie and Bhatia, 1998). Similarly, life expectancy and infant mortality rates are considered reliable indications of health when large populations are compared. We therefore continue, postulating that the measure we propose will corre-

late with syntactic differences as a whole, even if it does not measure them directly.

In fact we can be rather optimistic about using POS trigrams 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, §§ 2.2, 5.3, 4.1) demonstrates that this was common to theories in the 1980's (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). There is, on the other hand, consensus that the very strict lexicalism which Sells's work sketched must be relaxed in favor of "constructionalism" (Fillmore and Kay, 1999), but even in such theories syntactic heads have a privileged, albeit less dominant status.[1]

Let us further note that the focus on POS trigrams is poised to identify not only deviant syntactic uses, such as the one given as an example above, but also overuse and under-use of linguistic structure, whose importance is emphasized by researchers on second-language acquisition (Coseriu, 1970), (de Bot et al., 2005, A3,B3). 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 underuse) others. For example, Bot et al. (2005) suggest that non-transparent constructions are systematically avoided even by very good second-language learners).

2.1 Tagging

We tagged the material using Thorsten Brants's *Trigrams 'n Tags* (TnT) tagger, a hidden Markov model tagger which has performed at state-of-the-art levels in organized comparisons, achieving 96.7% correct on the material of the Penn Treebank (Brants, 2000).

Since our material is spoken English (see below), we trained the tagger on the spoken part of the *International Corpus of English* (ICE) from Great Britain, which consists of 500k words. This was suboptimal, as the material we wished to analyze was the English of Finnish emigrants to Australia, but we were unable to acquire sufficient Australian material.

We used the tagset of the TOSCA-ICE consisting of 270 tags (Garside et al., 1997), of which 75 were never instantiated in our material. In a sample of $1,000$ words we found that the tagger was correct for 87% of words, 74% of the bigrams, and 65% of the trigrams. As will be obvious in the presentation of the material (below), it is free conversation with pervasive foreign influence. We attribute the low tagging accuracy to the roughness of the material. It is clear that our procedure would improve in accuracy from a more accurate tagger, which would, in turn, allow application to smaller corpora.

We collect the material into a frequency vector containing the counts of $13,784$ different POS trigrams, one vector for each of the two sub-corpora which we describe below. We then ask whether the material in the one sub-corpus differs significantly from that in the other. We turn now to that topic.

3 Permutation Tests

There is no convenient test we can apply to check whether the differences between vectors containing $13,784$ elements are statistically significant, nor how significant the differences are. Fortunately, we may turn to permutation tests in this situation (Good, 1995), more specifically a permutation test using a Monte Carlo technique. Kessler (2001) contains an informal introduction for an application within linguistics.

The fundamental idea in a permutation test is very simple: we measure the difference between two sets in some convenient fashion, obtaining $\delta(A, B)$. We then extract two sets at random from $A \cup B$, calling these A_1, B_1, and we calculate the difference between these two in the same fashion, $\delta(A_1, B_1)$, recording the number of times $\delta(A_1, B_1) \geq \delta(A, B)$, i.e., how often two randomly selected subsets from the entire set of observations are at least as different as (usually more different than) the original sets were. If we repeat this process, say, $10,000$ times, then n, the number of times we obtain more extreme differences, 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, then the original division into A and B was likely to the degree of $p = n/10,000$. In more standard hypothesis-

[1] One referee suggested that one might test the association between POS trigram differences and head differences experimentally, and we find this suggestion sensible.

testing terms, this is the p-value with which we may reject (or retain) the null hypothesis that there is no relevant difference in the two sets.

We would like to guard against three dangers in our calculations. First, given the ease with which large corpora are obtained, we are uninterested in obtaining statistical significance through sheer corpus size. We aim therefore at obtaining a measure that is sensitive only to relative frequency, and not at all to absolute frequency (Agresti, 1996). Permutation tests effectively guard against this danger, if one takes care to judge samples of the same size within the permutations.

Second, we are mindful of a potential confounding factor, viz., the syntactical intra-dependence found within sentences (especially between adjoining POS trigrams). If we permuted n-grams, we might in part measure the internal coherence of the two initial sub-corpora, i.e., the coherence due to the fact that both sub-corpora use language conforming to the rules of English syntax. If we permuted n-grams, this coherence would be lost, and the measurement of difference would be affected. In the terminology of permutation statistics: the elements that are permuted must be reasonably independent. So we shall permute not n-grams, but rather entire sentences.

Third, the decision to permute sentences rather than n-grams exposes us to a confound due to systematically different sentence lengths. While the result of permuting elements in a Monte Carlo fashion always results in two sub-corpora that have the same number of elements as in the base-case, our problem is that the elements we permute are sentences, while what we measure are n-grams. Now if the original two sub-corpora differ substantially in average sentence length, then the result of the Monte Carlo "shuffling" will not be similar to the original split with respect to the number of n-grams involved. The original sub-corpus with longer sentences will therefore have many more n-grams in the base-case than in the random re-drawings from the combining corpora, at least on average. We address this danger systematically in the subsection below on within-permutation normalizations (§ 3.2).

We note a more subtle dependency we do not attempt to guard against. Some POS sequences (almost) only occur in relatively long sentences, e.g. the inversion that occurs in some conditionals *Were I in any doubt, I should not ....* Perhaps English subjunctives in general occur only in relatively long sentences. If this sort of structure occurs in one variety more frequently than in another, that is a genuine difference, but it might still be the reflection of the simpler difference in sentence length. One might then think that the second variety would show the same syntax if only it had longer sentences. As far as they are to be considered a problem in the first place, differences in syntax that are related to sentence length cannot be removed by (our) normalizations.

Permutation tests are a very suitable tool for finding significant syntactical differences, and for finding the POS trigrams that make a significant contribution to this difference.

3.1 Measuring Vector Differences

The choice of vector difference measure, e.g. cosine vs. χ^2, does not affect the proposed technique greatly, and alternative measures can be used straightforwardly. Accordingly, we have worked with both cosine and two measures inspired by the RECURRENCE (R) metric introduced by Kessler (Kessler, 2001, 157ff). Following Kessler, we also call our measures R and Rsq. The advantage of the R and Rsq metrics is that they are transparently interpretable as simple aggregates, meaning that one may easily see how much each trigram contributes to the overall corpus difference. We even used them to calculate a separate p-value per trigram.

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 ($\mathbf{c}$, $\mathbf{c}'$), and if i is the index of a POS trigram, R for each of these two vector cells is equal, as it is defined simply as $R = \sum_i |c_i - \overline{c_i}|$, with $\overline{c_i} = (c_i + c'_i)/2$. The Rsq measure attributes more weight to a few large differences than to many small ones, and it is calculated: $Rsq = \sum_i (c_i - \overline{c_i})^2$, with $\overline{c_i}$ being the same as above (for R).

3.2 Within-Permutation Normalization

Each measurement of difference—whether the difference is between the original two samples or between two samples which arise through permutations—is taken over the collection of POS trigram frequencies once these have been normalized. We describe first the normalization that is required to cope with differences in sentence length

which we call WITHIN-PERMUTATION NORMALIZATION, as it is applied within each permutation.

In case sub-corpora differ in sentence length, they will automatically differ in the number of n-grams across permutations as well. Our Monte Carlo choice of alternatives does not change the relative number of sentences across permutations, but the number of POS trigrams in the groups will vary if no normalization is applied. Longer sentences give rise to larger numbers of POS trigrams per sentence, and therefore per sub-corpora. Applying the within-permutation normalization one or more times ensures that this does not infect the measurement of difference.

Protecting the measurement from sensitivity to differing numbers of POS trigrams per sentence is for us sufficient reason to normalize, but we also normalize in order to facilitate interpretation. We return to this below, in the definition of the rescaled vectors $\mathbf{s^y}, \mathbf{s^o}$.

We thus collect from the tagger a sequence of counts c_i of tag trigrams for each sample. We treat only the case of comparing two samples here, which we shall refer to as young (y) and old (o) for reasons which will become clear in the following section. We shall keep track of the sum-per-tag trigram as well, summing over the two sub-corpora.

$$\begin{array}{rcll} \mathbf{c^y} & = & < c_1^y, c_2^y, ..., c_n^y > & N^y = \sum_{i=1}^n c_i^y \\ +\mathbf{c^o} & = & < c_1^o, c_2^o, ..., c_n^o > & N^o = \sum_{i=1}^n c_i^o \\ \hline \mathbf{c} & = & < c_1, c_2, ..., c_n > & N(= N^y + N^o) \\ & & & = \sum_{i=1}^n c_i \end{array}$$

As a first step in normalization, we work with vectors holding the relative frequency fractions per group:

$$\begin{aligned} \mathbf{f^y} &= < ..., f_i^y (= c_i^y / N^y), ... > \\ \mathbf{f^o} &= < ..., f_i^o (= c_i^o / N^o), ... > \end{aligned}$$

We note that $\sum_{i=1}^n f_i^y = \sum_{i=1}^n f_i^o = 1$.

We then compute the relative proportions per trigram, comparing now across the groups. This prepares for the step which redistributes the raw trigram counts to compensate for differences in sentence length.

$$\begin{aligned} \mathbf{p^y} &= < ..., p_i^y (= f_i^y / (f_i^y + f_i^o)), ... > \\ \mathbf{p^o} &= < ..., p_i^o (= f_i^o / (f_i^y + f_i^o)), ... > \end{aligned}$$

We might also define a sum of $\mathbf{p^y} + \mathbf{p^o}$:

$$\mathbf{p} = < ..., p_i (= (p_i^o + p_i^y) = 1), ... >$$

We do not actually use $\mathbf{p}$ below, only $\mathbf{p^y}$ and $\mathbf{p^o}$, but we mention it for the sake of the check it allows that $p_i^y + p_i^o = 1, \forall i$.

We then re-introduce the raw frequencies per category to obtain the normalized, redistributed counts $\mathbf{C_n^y}, \mathbf{C_n^o}$. Note that we use the total count of the trigram in both samples to redistribute (thus redistributing these counts based on the trigram totals in both samples):

$$\begin{aligned} \mathbf{C_n^y} &= < ..., p_i^y \cdot c_i, ... > \\ \mathbf{C_n^o} &= < ..., p_i^o \cdot c_i, ... > \end{aligned}$$

Up to this point the normalization has corrected for differences in sentence length, or to be more precise, for differences in the numbers of n-grams which may appear as a result of permuting sentences. For larger numbers of trigrams the situation will become: $N^y = \sum_{i=1}^n c_i^y \approx \sum_{i=1}^n \mathbf{C_i^y}$ so that we have effectively neutralized the increase or decrease in the number of n-grams which might have arisen due to sentence length. Without this normalization a skew in sentence length in the base case would cause changed, in the worst case increased, and perhaps even extreme, significance. During random permutation, where longer sentences will tend to be distributed more evenly between the sub-corpora, a disproportionately larger number of n-grams would be found in the sub-corpus corresponding to the base corpus with shorter sentences. We have now normalized so that that effect will no longer appear.

We illustrate the normalizations up to this point in Table 1. We see already that the overall effect is to shift mass to the smaller sample. Notice also that if we were to define $\mathbf{C} = \mathbf{C^y} + \mathbf{C^o}$, then $\mathbf{C} = \mathbf{c}$, since $\mathbf{C^y}$ and $\mathbf{C^o}$ are a redistribution of $\mathbf{c}$ using p^y and p^o, whose sum p is 1 under all circumstances, as was noted above. At the same time $\mathbf{c^y} \neq \mathbf{C^y}$ and $\mathbf{c^o} \neq \mathbf{C^o}$ (if there were differences in sentence lengths). The values obtained at this point may be measured by the vector comparison measure (cosine or $R(sq)$).

We use this redistributing normalization instead of just the relative frequency because using relative frequency would cause trigrams occurring mainly and frequently in the short-sentence group to become extremely significant. This is especially

	Group y		Group o		Group y'		Group o'	
	T1	T2	T1	T2	T1	T2	T1	T2
counts $\mathbf{c}$	**15**	**10**	90	10	10	10	17	0
rel. freq. $\mathbf{f}$	0.6	0.4	0.9	0.1	0.5	0.5	1	0
norm. prop. $\mathbf{p}$	0.4	0.8	0.6	0.2	0.33	1	0.67	0
trigram $\mathbf{c}_i$	105	20	105	20	27	10	27	10
redistrib. $\mathbf{C}$	42	16	63	4	9	10	18	0

Table 1: Two examples of the normalizations applied before each measurement of vector difference. On the left groups y and o are compared on the basis of the two trigrams $T1$ and $T2$. The counts are shown in the first row, then relative frequencies (within the group), normalized relative proportions, and finally redistributed normalized counts. The two numbers in boldface in the 'count' line are compared to calculate the underlined relative frequency (on the left) in the 'relative frequency' line (in general counts are compared within groups to obtain relative frequencies). Next, the two underlined fractions of the 'relative frequency' row are compared to obtain the corresponding fractions (immediately below) of the 'normalized proportions' row. Thus relative frequencies are compared across groups (sub-corpora) to obtain the relative proportions. The trigram count row shows the counts per trigram type, and the 'redistributed' row is simply the product of the last two. The second example (on the right) demonstrates that missing data finds no compensation in this procedure (although we might experiment with smoothing in the future).

distorting if one calculates the per trigram type p-value (R or Rsq for a single i).

The normalization does not eliminate all the irrelevant effects of differing sentence lengths. To obtain further precision we iterate the steps above a few times, re-applying the normalization to its own output. We are motivated to iterate the procedure for the following reason. If a trigram is relatively more frequent in the smaller sub-corpus, it must then also be relatively less frequent within the entire corpus (less frequent within the two sub-corpora together), so there is less frequency mass to re-distribute for these trigrams than for trigrams that are relatively more frequent in the larger sub-corpus (those will be more frequent within the entire corpus). A special case of this are n-grams that occur only in one sub-corpus. If they occur only in the larger sub-corpus then their mass will never be re-distributed in the direction of the smaller sub-corpus, since zero-frequencies within one sub-corpus will always result in zero relative weight (in the current set-up).[2] This means that after normalization the larger sub-corpus will always still be a bit larger than the smaller one. After one normalization the effect of these factors is small, but we can reduce it yet further by iterating the normalization. This is worthwhile since we wish to be certain. After five iterations the relative size-difference between our normalized sub-corpora is less than 0.1% for trigrams of the full ICE-tagset (and even a thousand times smaller for the reduced tagset). We regard this as small enough to effectively eliminate corpus size differences as potential problems.

For the purposes of interpretation we also scale everything down so that the average redistributed count is 1. We do this by dividing each C_i^y, C_i^o by $N/2n$, where N is the total count of all trigrams and n is the number of trigram categories being counted. Note that $N/2n$ is the average count of a given trigram in one of the groups.

$$\begin{aligned} \mathbf{s^y} &= \mathbf{c^y} \cdot 2n/N &= &< ..., C_i^y \cdot 2n/N, ... > \\ \mathbf{s^o} &= \mathbf{c^o} \cdot 2n/N &= &< ..., C_i^o \cdot 2n/N, ... > \end{aligned}$$

These values might just as well be submitted to the vector comparison measure since they are just linear transformations of the redistributed $\mathbf{C}$ values. The scaling expresses the trigram count as a value with respect to the total $2n$ of counts involved in the comparison, and, since $\sum_{i=1}^n c_i^y + \sum_{i=1}^n c_i^o = N$, $\sum_{i=1}^n s_i^y + \sum_{i=1}^n s_i^o = 2n$. As there are n sorts of trigrams being compared in two groups, it is clear that the average value in these last vectors will be 1.

Similarly, this normalized value will be higher than 1 for trigrams that are more frequent than average. Now if we sort the trigrams by frequency—

[2] Alternatively, we might have explored a Good-Turing estimation of unseen items (Manning and Schütze, 1999, p. 212).

or more precisely, by the weight that they have within the total $R(sq)$ value, so by their per trigram $R(sq)$ value—we get a listing of the POS trigrams that distinguish the groups most sharply. This list can be made even more telling by adding the raw frequency and a per-trigram p-value. It allows us to directly see significant under and overuse of POS trigrams, and thereby of syntax.

3.3 Between-Permutations Normalization

The purpose of this normalization is the identification of n-gram types which are typical in the two original sub-corpora. It is applied after comparing all the results of all the Monte Carlo re-shufflings.

The BETWEEN-PERMUTATIONS NORMALIZATION is similar to the last step of the within-permutation normalization, except that the linear transformation is applied across permutations, instead of across groups (sub-corpora): for each POS trigram type i in each group (sub-corpora) $g \in \{o, y\}$, the redistributed count $\mathbf{C_i^g}$ is divided by the average redistributed count for that type in that group (across all permutations)$\overline{\mathbf{C_i^g}}$. Note that the average redistributed count is $c_i/2$ for large numbers of permutations. The values thus normalized will be 1 on average across permutations.

Trigrams with large average counts between permutations are those with high frequencies in the original sub-corpora, and these contribute most heavily toward statistical significance. The normalization under discussion strips away the role of frequency, allowing us to see which POS trigrams are most (a)typical for a group. We note additionally that this normalization is useful only together with information on frequency (or statistical significance). Infrequent trigrams are especially likely to have high values with respect to $\overline{\mathbf{C_i^g}}$. For example a trigram occurring only once, in one sub-corpus, gets the maximum value of $1/0.5 = 2$ (as it is indeed very typical for this sub-corpus), while with a count of one it clearly cannot be statistically significant (moving between equally sized sub-corpora with a chance of 50 % during permutations). So it's best to calculate this normalization together with the per trigram p-values.

4 A Test Case

We tested this procedure on data transcribed from free interviews with Finnish emigrants to Australia. The emigrants were farmers and skilled or semi-skilled working class Finns who left Finland in the 1960's at the age of 25-40 years old, some with children. Greg Watson of Joensuu University interviewed these people between 1995 and 1998, publishing about his corpus in *ICAME* 20, 1996 (Watson). He included both interviews with those who emigrated as adults (at seventeen years or older) and those who emigrated as children (before their seventeenth birthday). There are sixty conversations with adult-age emigrants and thirty with those who emigrated as children, totaling 305,000 words of relatively free conversation.

It is well established in the literature on second-language learning that the language of people who learned the second language as children is superior to that of adult learners. We will test our idea about measuring syntactic differences by applying the measure to the two samples language from adult vs. child emigrants. The issue is not remarkable, but it allows us to verify whether the measure is functioning.

4.1 Results

The two sub-corpora had 221,000 words for the older group and 84,000 words for the younger group, respectively. The sentences of the childhood immigrants were indeed substantially longer (27.1 tokens) than those of the older immigrants (16.3 tokens). So the within-permutation normalization was definitely needed in this case. The groups clearly differed in the distribution of POS trigrams they contain ($p < 0.001$). This means that the difference between the original two sub-corpora was in the largest 0.1% of the Monte Carlo permutations.

In addition we find genuinely deviant syntax patterns if we inspect the trigrams most responsible for the difference between the two sub-corpora.

it PRO	's COP	low ADJ	tax N/COM	in PREP	here ADV
and CONJ	I PRO	was COP	professional ADJ	fisherman N/COM	

Both COP-ADJ-N/COM and N/COM-PREP-ADV accounted for a substantial degree of aggregate syntactic difference. The first pattern normally corresponds to an error, as it does in the two (!) examples of it above (there is a separate tag for plural and mass nouns). These are cases where English normally requires an article.

Since Finnish has no articles, these are clear cases of transfer, i.e., the (incorrect) imposition of the first language's structure on a second language. The N/COM-PREP-ADV pattern (corresponding to the use of *in here*) is also worth noting, as it falls into the class of expressions which is not absolutely in error (*The material is in here*), but it is clearly being overused in the example above. Presumably this is a case of hypercorrection from Finnish, a language without prepositions. We conclude from this experiment that the procedure is promising.

On the other hand there were also problems, perhaps most seriously with the use of the tags denoting pauses and hesitations, where we found that the tag trigrams most responsible for the deviant measures in the corpora involved disfluencies of one sort or another. These tended to occur more frequently in the speech of the older emigrants. With the pauses removed (hesitations still in place) a list of the ten most frequent, significant trigrams for the older group is shown. Two random examples from the corpus are given for each in Table 2.

We suspect additionally that the low accuracy rate of the tagger when applied to this material also stems from the large number of disfluencies.

5 Conclusions and Prospects

Weinreich (1953) regretted that there was no way to "measure or characterize the total impact one language on another in the speech of bilinguals," (p. 63) and speculated that there could not be. This paper has proposed a way of going beyond counts of individual phenomena to a measure of aggregate syntactic difference. The technique may be implemented effectively, and its results are subject to statistical analysis using permutation statistics.

The technique proposed follow Aarts and Granger (1998) in using part-of-speech trigrams. 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 go beyond Aarts and Granger in Showing how entire histograms of POS trigrams may be used to characterize aggregate syntactic distance, in particular by showing how this can be analyzed.

We fall short of Weinreich's goal of assaying "total impact" in that we focus on syntax, but we

1	roadworks	and	uh
	hill	and	ah
	N	CONJUNC	INTERJEC
2	I	reckon	it
	that	take	lot
	PRON	V	PRON
3	enjoy	to	taking
	my	machine	break
	INTERJEC	PRON	V
4	but	that	's
	that	I	clean
	CONJUNC	PRON	V
5	I	'm	uh
	it	's	uh
	PRON	V	INTERJEC
6	now	what	what
	changing	but	some
	CONJUNC	INTERJEC	PRON
7	said	it	's
	all	everybody	has
	PRON	PRON	V
8	bought	that	car
	lead	glass	windows
	V	PRON	N
9	that	was	different
	I	was	fit
	PRON	V	ADJ
10	Oh	lake	lake
	uh	money	production
	INTERJEC	N	N

Table 2: The most significant and most frequent trigrams that were typical for the speech of the group of older Finnish emigrants to Australia compared to the speech of those who emigrated before their 17th birthday. The tag trigrams indicating pauses were removed before comparing the corpora, as these appear to dominate the differences. The examples illustrating the trigrams were chosen at random, and we note that the examples of the third sort of trigram involved tagging errors in the first and second elements of the trigram, and that other errors are noticeable at the seventh and eight positions in the list (where 'said' and 'glass' are marked as pronouns). We reserve the linguistic interpretation of the error patterns for future work, but we note that we will also want to filter interjections before drawing definite conclusions.

take a large step in this direction by showing how to aggregate and test for significance, using the sorts of counts he worked with.

The software implementing the permutation test, including the normalizations, is available freely at `http://en.logilogi.org/HomE/WyboWiersma/FiAuImEnRe`. It is developed to allow easy generalization to more than two sub-corpora and longer n-grams.

Several further steps would be useful. We should like to repeat the analysis here, eliminating the effect of hesitation tags, etc. Second, we should like to experiment systematically with the inclusion of n-grams for $n > 3$; to-date we have experimented with this, but not systematically enough. Third, we would like to test the analysis on other cases of putative syntactic differences, and in particular in cases where tagging accuracy might be less an issue.

Acknowledgments

We are grateful to Lisa Lena Opas-Hänninen, Pekka Hirvonen and Timo Lauttamus 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*, Aug. 2005; the Sonderforschungsbereich 441, "Linguistic Data Structures", Tübingen in Jan. 2006; and the Seminar on Methodology and Statistics in Linguistic Research, University of Groningen, Spring, 2006, and especially Livi Ruffle, for useful comments and discussion. Finally, two referees for the 2006 ACL/COLING workshop on "Linguistic Distances" also commented usefully.

References

Jan Aarts and Sylviane Granger. 1998. Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In Sylviane Granger, editor, *Learner English on Computer*, pages 132–141. Longman, London.

Alan Agresti. 1996. *An Introduction to Categorical Data Analysis*. Wiley, New York.

Thorsten Brants. 2000. TnT — a statistical part of speech tagger. In *6th Applied Natural Language Processing Conference*, pages 224–231, Seattle. ACL.

Eugenio Coseriu. 1970. *Probleme der kontrastiven Grammatik*. Schwann, Düsseldorf.

Kees de Bot, Wander Lowie, and Marjolijn Verspoor. 2005. *Second Language Acquisition: An Advanced Resource Book*. Routledge, London.

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

Roger Garside, Geoffrey Leech, and Tony McEmery. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London/New York.

Phillip Good. 1995. *Permutation Tests*. Springer, New York. 2nd, corr. ed.

1949. *The Principles of the International Phonetic Association*. International Phonetics Association, London, 1949.

Brett Kessler. 2001. *The Significance of Word Lists*. CSLI Press, Stanford.

Chris Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

Shana Poplack and David Sankoff. 1984. Borrowing: the synchrony of integration. *Linguistics*, 22:99–135.

Shana Poplack, David Sankoff, and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26:47–104.

William C. Ritchie and Tej K. Bhatia, editors. 1998. *Handbook of Child Language Acquisition*. Academic, San Diego.

Peter Sells. 1982. *Lectures on Contemporary Syntactic Theories*. CSLI, Stanford.

Sarah Thomason and Terrence Kaufmann. 1988. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley.

Frans van Coetsem. 1988. *Loan Phonology and the Two Transfer Types in Language Contact*. Publications in Language Sciences. Foris Publications, Dordrecht.

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

Uriel Weinreich. 1953. *Languages in Contact*. Mouton, The Hague. (page numbers from 2nd ed. 1968).