Measuring Syntactic Difference in British English

Nathan C. Sanders
Department of Linguistics
Indiana University
Bloomington, IN 47405, USA
ncsander@indiana.edu

Abstract

Recent work by Nerbonne and Wiersma (2006) has provided a foundation for measuring syntactic differences between corpora. It uses part-of-speech trigrams as an approximation to syntactic structure, comparing the trigrams of two corpora for statistically significant differences.

This paper extends the method and its application. It extends the method by using leaf-path ancestors of Sampson (2000) instead of trigrams, which capture internal syntactic structure—every leaf in a parse tree records the path back to the root.

The corpus used for testing is the International Corpus of English, Great Britain (Nelson et al., 2002), which contains syntactically annotated speech of Great Britain. The speakers are grouped into geographical regions based on place of birth. This is different in both nature and number than previous experiments, which found differences between two groups of Norwegian L2 learners of English. We show that dialectal variation in eleven British regions from the ICE-GB is detectable by our algorithm, using both leaf-ancestor paths and trigrams.

1 Introduction

In the measurement of linguistic distance, older work such as Séguy (1973) was able to measure distance in most areas of linguistics, such as phonology, morphology, and syntax. The features used for comparison were hand-picked based on linguistic knowledge of the area being surveyed. These features, while probably lacking in completeness of coverage, certainly allowed a rough comparison of distance in all linguistic domains. In contrast, computational methods have focused on a single area of language. For example, a method for determining phonetic distance is given by Heeringa (2004). Heeringa and others have also done related work on phonological distance in Nerbonne and Heeringa (1997) and Gooskens and Heeringa (2004). A measure of syntactic distance is the obvious next step: Nerbonne and Wiersma (2006) provide one such method. This method approximates internal syntactic structure using vectors of part-of-speech trigrams. The trigram types can then be compared for statistically significant differences using a permutation test.

This study can be extended in a few ways. First, the trigram approximation works well, but it does not necessarily capture all the information of syntactic structure such as long-distance movement. Second, the experiments did not test data for geographical dialect variation, but compared two generations of Norwegian L2 learners of English, with differences between ages of initial acquisition.

We address these areas by using the syntactically annotated speech section of the International Corpus of English, Great Britain (ICE-GB) (Nelson et al., 2002), which provides a corpus with full syntactic annotations, one that can be divided into groups for comparison. The sentences of the corpus, being represented as parse trees rather than a vector of POS tags, are converted into a vector of leaf-ancestor paths, which were developed by Sampson (2000) to aid in parser evaluation by providing a way to compare gold-standard trees with parser output trees.

In this way, each sentence produces its own vec-

Proceedings of the ACL 2007 Student Research Workshop, pages 1–6,
Prague, June 2007.

tor of leaf-ancestor paths. Fortunately, the permutation test used by Nerbonne and Wiersma (2006) is already designed to normalize the effects of differing sentence length when combining POS trigrams into a single vector per region. The only change needed is the substitution of leaf-ancestor paths for trigrams.

The speakers in the ICE-GB are divided by place of birth into geographical regions of England based on the nine Government Office Regions, plus Scotland and Wales. The average region contains a little over 4,000 sentences and 40,000 words. This is less than the size of the Norwegian corpora, and leaf-ancestor paths are more complex than trigrams, meaning that the amount of data required for obtaining significance should increase. Testing on smaller corpora should quickly show whether corpus size can be reduced without losing the ability to detect differences.

Experimental results show that differences can be detected among the larger regions: as should be expected with a method that measures statistical significance, larger corpora allow easier detection of significance. The limit seems to be around 250,000 words for leaf-ancestor paths, and 100,000 words for POS trigrams, but more careful tests are needed to verify this. Comparisons to judgments of dialectologists have not yet been made. The comparison is difficult because of the difference in methodology and amount of detail in reporting. Dialectology tends to collect data from a few informants at each location and to provide a more complex account of relationship than the like/unlike judgments provided by permutation tests.

2 Methods

The methods used to implement the syntactic difference test come from two sources. The primary source is the syntactic comparison of Nerbonne and Wiersma (2006), which uses a permutation test, explained in Good (1995) and in particular for linguistic purposes in Kessler (2001). Their permutation test collects POS trigrams from a random subcorpus of sentences sampled from the combined corpora. The trigram frequencies are normalized to neutralize the effects of sentence length, then compared to the trigram frequencies of the complete corpora.

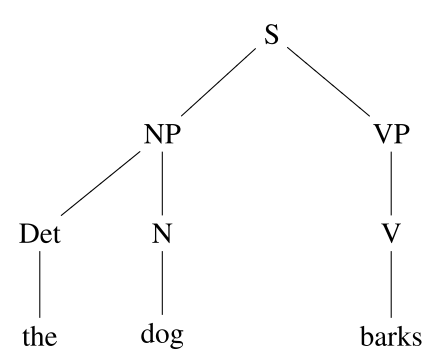
The principal difference between the work of Nerbonne and Wiersma (2006) and ours is the use of leaf-ancestor paths. Leaf-ancestor paths were developed by Sampson (2000) for estimating parser performance by providing a measure of similarity of two trees, in particular a gold-standard tree and a machine-parsed tree. This distance is not used for our method, since for our purposes, it is enough that leaf-ancestor paths represent syntactic information, such as upper-level tree structure, more explicitly than trigrams.

The permutation test used by Nerbonne and Wiersma (2006) is independent of the type of item whose frequency is measured, treating the items as atomic symbols. Therefore, leaf-ancestor paths should do just as well as trigrams as long as they do not introduce any additional constraints on how they are generated from the corpus. Fortunately, this is not the case; Nerbonne and Wiersma (2006) generate $N - 2$ POS trigrams from each sentence of length N; we generate N leaf-ancestor paths from each parsed sentence in the corpus. Normalization is needed to account for the frequency differences caused by sentence length variation; it is presented below. Since the same number (minus two) of trigrams and leaf-ancestor paths are generated for each sentence the same normalization can be used for both methods.

2.1 Leaf-Ancestor Paths

Sampson's leaf-ancestor paths represent syntactic structure by aggregating nodes starting from each leaf and proceeding up to the root—for our experiment, the leaves are parts of speech. This maintains constant input from the lexical items of the sentence, while giving the parse tree some weight in the representation.

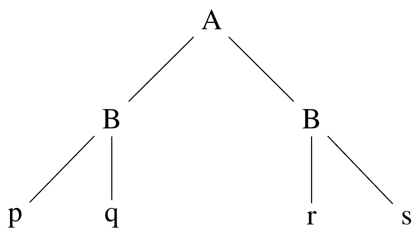
For example, the parse tree



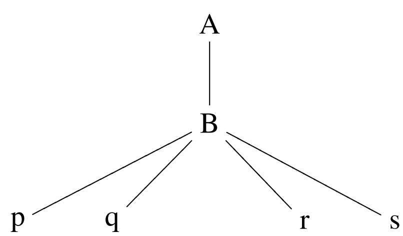
creates the following leaf-ancestor paths:

- S-NP-Det-The
- S-NP-N-dog
- S-VP-V-barks

There is one path for each word, and the root appears in all four. However, there can be ambiguities if some node happens to have identical siblings. Sampson gives the example of the two trees



and



which would both produce

- A-B-p
- A-B-q
- A-B-r
- A-B-s

There is no way to tell from the paths which leaves belong to which B node in the first tree, and there is no way to tell the paths of the two trees apart despite their different structure. To avoid this ambiguity, Sampson uses a bracketing system; brackets are inserted at appropriate points to produce

- [A-B-p
- A-B]-q
- A-[B-r
- A]-B-s

and

- [A-B-p
- A-B-q
- A-B-r
- A]-B-s

Left and right brackets are inserted: at most one in every path. A left bracket is inserted in a path containing a leaf that is a leftmost sibling and a right bracket is inserted in a path containing a leaf that is a rightmost sibling. The bracket is inserted at the highest node for which the leaf is leftmost or rightmost.

It is a good exercise to derive the bracketing of the previous two trees in detail. In the first tree, with two B siblings, the first path is A-B-p. Since p is a leftmost child, a left bracket must be inserted, at the root in this case. The resulting path is [A-B-p. The next leaf, q, is rightmost, so a right bracket must be inserted. The highest node for which it is rightmost is B, because the rightmost leaf of A is s. The resulting path is A-B]-q. Contrast this with the path for q in the second tree; here q is not rightmost, so no bracket is inserted and the resulting path is A-B-q. r is in almost the same position as q, but reversed: it is the leftmost, and the right B is the highest node for which it is the leftmost, producing A-[B-r. Finally, since s is the rightmost leaf of the entire sentence, the right bracket appears after A: A]-B-s.

At this point, the alert reader will have noticed that both a left bracket and right bracket can be inserted for a leaf with no siblings since it is both leftmost and rightmost. That is, a path with two brackets on the same node could be produced: A-[B]-c. Because of this redundancy, single children are excluded by the bracket markup algorithm. There is still no ambiguity between two single leaves and a single node with two leaves because only the second case will receive brackets.

2.2 Permutation Significance Test

With the paths of each sentence generated from the corpus, then sorted by type into vectors, we now try to determine whether the paths of one region occur in significantly different numbers from the paths of another region. To do this, we calculate some measure to characterize the difference between two vectors as a single number. Kessler (2001) creates a simple measure called the RECURRENCE metric (R hereafter), which is simply the sum of absolute differences of all path token counts c_{ai} from the first corpus A and c_{bi} from the second corpus B.

$$R = \Sigma_i |c_{ai} - \bar{c}_i| \text{ where } \bar{c}_i = \frac{c_{ai} + c_{bi}}{2}$$

However, to find out if the value of R is significant, we must use a permutation test with a Monte Carlo technique described by Good (1995), following closely the same usage by Nerbonne and Wiersma (2006). The intuition behind the technique is to compare the R of the two corpora with the R of two random subsets of the combined corpora. If the random subsets' Rs are greater than the R of the two actual corpora more than p percent of the time, then we can reject the null hypothesis that the two were are actually drawn from the same corpus: that is, we can assume that the two corpora are different.

However, before the R values can be compared, the path counts in the random subsets must be normalized since not all paths will occur in every subset, and average sentence length will differ, causing relative path frequency to vary. There are two normalizations that must occur: normalization with respect to sentence length, and normalization with respect to other paths within a subset.

The first stage of normalization normalizes the counts for each path within the pair of vectors a and b. The purpose is to neutralize the difference in sentence length, in which longer sentences with more words cause paths to be relatively less frequent. Each count is converted to a frequency f

$$f = \frac{c}{N}$$

where c is either c_{ai} or c_{bi} from above and N is the length of the containing vector a or b. This produces two frequencies, f_{ai} and f_{bi}.Then the frequency is scaled back up to a redistributed count by the equation

$$\forall j \in a, b : c'_{ji} = \frac{f_{ji}(c_{ai} + c_{bi})}{f_{ai} + f_{bi}}$$

This will redistribute the total of a pair from a and b based on their relative frequencies. In other words, the total of each path type $c_{ai} + c_{bi}$ will remain the same, but the values of c_{ai} and c_{bi} will be balanced by their frequency within their respective vectors.

For example, assume that the two corpora have 10 sentences each, with a corpus a with only 40 words and another, b, with 100 words. This results in $N_a = 40$ and $N_b = 100$. Assume also that there is a path i that occurs in both: $c_{ai} = 8$ in a and $c_{bi} = 10$ in b. This means that the relative frequencies are $f_{ai} = 8/40 = 0.2$ and $f_{bi} = 10/100 = 0.1$. The first normalization will redistribute the total count (18) according to relative size of the frequencies. So

$$c'_{ai} = \frac{0.2(18)}{0.2 + 0.1} = 3.6/0.3 = 12$$

and

$$c'_{bi} = \frac{0.1(18)}{0.2 + 0.1} = 1.8/0.3 = 6$$

Now that 8 has been scaled to 12 and 10 to 6, the effect of sentence length has been neutralized. This reflects the intuition that something that occurs 8 of 40 times is more important than something that occurs 10 of 100 times.

The second normalization normalizes all values in both permutations with respect to each other. This is simple: find the average number of times each path appears, then divide each scaled count by it. This produces numbers whose average is 1.0 and whose values are multiples of the amount that they are greater than the average. The average path count is $N/2n$, where N is the number of path tokens in both the permutations and n is the number of path types. Division by two is necessary since we are multiplying counts from a single permutation by token counts from both permutations. Each type entry in the vector now becomes

$$\forall j \in a, b : s_{ji} = \frac{2nc'_{ji}}{N}$$

Starting from the previous example, this second normalization first finds the average. Assuming 5 unique paths (types) for a and 30 for b gives

$$n = 5 + 30 = 35$$

and

$$N = N_a + N_b = 40 + 100 = 140$$

Therefore, the average path type has $140/2(35) = 2$ tokens in a and b respectively. Dividing c'_{ai} and c'_{bi} by this average gives $s_{ai} = 6$ and $s_{bi} = 3$. In other words, s_{ai} has 6 times more tokens than the average path type.

Region	sentences	words
East England	855	10471
East Midlands	1944	16924
London	24836	244341
Northwest England	3219	27070
Northeast England	1012	10199
Scotland	2886	27198
Southeast England	11090	88915
Southwest England	939	7107
West Midlands	960	12670
Wales	2338	27911
Yorkshire	1427	19092

Table 1: Subcorpus size

Region	Significantly different ($p < 0.05$)
London	East Midlands, NW England SE England, Scotland
SE England	Scotland

Table 2: Significant differences, leaf-ancestor paths

Region	Significantly different ($p < 0.05$)
London	East Midlands, NW England, NE England, SE England, Scotland, Wales
SE England	London, East Midlands, NW England, Scotland
Scotland	London, SE England, Yorkshire

Table 3: Significant differences, POS trigrams

3 Experiment and Results

The experiment was run on the syntactically annotated part of the International Corpus of English, Great Britain corpus (ICE-GB). The syntactic annotation labels terminals with one of twenty parts of speech and internal nodes with a category and a function marker. Therefore, the leaf-ancestor paths each started at the root of the sentence and ended with a part of speech. For comparison to the experiment conducted by Nerbonne and Wiersma (2006), the experiment was also run with POS trigrams. Finally, a control experiment was conducted by comparing two permutations from the same corpus and ensuring that they were not significantly different.

ICE-GB reports the place of birth of each speaker, which is the best available approximation to which dialect a speaker uses. As a simple, objective partitioning, the speakers were divided into 11 geographical regions based on the 9 Government Office Regions of England with Wales and Scotland added as single regions. Some speakers had to be thrown out at this point because they lacked brithplace information or were born outside the UK. Each region varied in size; however, the average number of sentences per corpus was 4682, with an average of 44,726 words per corpus (see table 1). Thus, the average sentence length was 9.55 words. The average corpus was smaller than the Norwegian L2 English corpora of Nerbonne and Wiersma (2006), which had two groups, one with 221,000 words and the other with 84,000.

Significant differences (at $p < 0.05$) were found when comparing the largest regions, but no significant differences were found when comparing small regions to other small regions. The significant differences found are given in table 2 and 3. It seems that summed corpus size must reach a certain threshold before differences can be observed reliably: about 250,000 words for leaf-ancestor paths and 100,000 for trigrams. There are exceptions in both directions; the total size of London compared to Wales is larger than the size of London compared to the East Midlands, but the former is not statistically different. On the other hand, the total size of Southeast England compared to Scotland is only half of the other significantly different comparisons; this difference may be a result of more extreme syntactic differences than the other areas. Finally, it is interesting to note that the summed Norwegian corpus size is around 305,000 words, which is about three times the size needed for significance as estimated from the ICE-GB data.

4 Discussion

Our work extends that of Nerbonne and Wiersma (2006) in a number of ways. We have shown that an alternate method of representing syntax still allows the permutation test to find significant differences between corpora. In addition, we have shown differences between corpora divided by geographical area rather than language proficiency, with many more corpora than before. Finally, we have shown that the size of the corpus can be reduced somewhat and still obtain significant results.

Furthermore, we also have shown that both leaf-ancestor paths and POS trigrams give similar results, although the more complex paths require more data.

However, there are a number of directions that this experiment should be extended. A comparison that divides the speakers into traditional British dialect areas is needed to see if the same differences can be detected. This is very likely, because corpus divisions that better reflect reality have a better chance of achieving a significant difference.

In fact, even though leaf-ancestor paths should provide finer distinctions than trigrams and thus require more data for detectable significance, the regional corpora presented here were smaller than the Norwegian speakers' corpora in Nerbonne and Wiersma (2006) by up to a factor of 10. This raises the question of a lower limit on corpus size. Our experiment suggests that the two corpora must have at least 250,000 words, although we suspect that better divisions will allow smaller corpus sizes.

While we are reducing corpus size, we might as well compare the increasing numbers of smaller and smaller corpora in an advantageous order. It should be possible to cluster corpora by the point at which they fail to achieve a significant difference when split from a larger corpus. In this way, regions could be grouped by their detectable boundaries, not a priori distinctions based on geography or existing knowledge of dialect boundaries.

Of course this indirect method would not be needed if one had a direct method for clustering speakers, by distance or other measure. Development of such a method is worthwhile research for the future.

References

Phillip Good. 1995. *Permutation Tests*. Springer, New York.

Charlotte S. Gooskens and Wilbert J. Heeringa. 2004. Perceptive evaluations of levenshtein dialect distance measurements using norwegian dialect data. *Language Variation and Change*, 16(3):189–207.

Wilbert J. Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Doctoral dissertation, University of Groningen.

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

Gerald Nelson, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: working with the British component of the International Corpus of English*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

John Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In John Coleman, editor, *Workshop on Computational Phonology*, pages 11–18, Madrid. Special Interest Group of the Assocation for Computational Linguistics.

John Nerbonne and Wybo Wiersma. 2006. A measure of aggregate syntactic distance. In John Nerbonne and Erhard Hinrichs, editors, *Linguistic Distances*, pages 82–90, Sydney, July. International Committee on Computational Linguistics and the Assocation for Computational Linguistics.

Geoffrey Sampson. 2000. A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics*, 5(1):53–68, August.

Jean Séguy. 1973. La dialectometrie dans l'atlas linguistique de la gascogne. *Revue de linguistique romane*, 37:1–24.