Found in Translation: Computational Discovery of Translation Effects

Carl Vogel, Gerard Lynch, Erwan Moreau, Liliana Mamani Sanchez, Phil Ritchie

Found in Translation: Computational Discovery of Translation Effects

Carl Vogel, Ger Lynch, Erwan Moreau,
Liliana Mamani Sanchez and Phil Ritchie
Computational Linguistics Group,
School of Computer Science and Statistics,
Trinity College, Dublin 2, Ireland
email: vogel@tcd.ie

May 3, 2013

Abstract

We describe translation effects that have been studied in the the automated text classification literature. We expand on a point within this research space, quality effects, with our own work in this area. We present an efficient method for evaluating text quality on the basis of reference texts. The method, which is general to text classification problems more widely construed, is related to the background literature and argued to be effective on the strength of the fact that it enables quality checking of amounts of text that exceed what is humanly feasible to verify. The method partially automates the process: in processing the entirety of a translated corpus being probed, it ranks items for stylistic conformity with a reference corpus, and the least conforming ranks are indicated as the items most likely to require human intervention.

1 Introduction

This article describes approaches to discovery of translation effects using automated methods. Translation effects are those features of texts which make visible that they are translations, independent of information about provenance that might happen to be available. The effects may be sufficient to reveal who the translator is, human or otherwise, or what the source language was for the text. The effects may also be second-order effects in which translations are adequate in themselves, but not in relation to an intended reference style. This is a point on a cline towards being a translation into the wrong target language. The problem setting for automatic analysis as we explore it here is one in which the volume of text to assess is too great to manually inspect an entire translated corpus. This could be the situation faced by a high-volume commercial translation service provider. For many purposes in such a scenario, one may rely on the

fact that trusted translators do excellent work, and as a result put to third-party quality assurance verifiers only samples from within any one translation project. This creates a question of reliable sampling. The method of analysis that we report on subjects the entire translation project to automated scrutiny in order to identify the subset that is most likely to benefit from manual analysis.

The method of analysis reported on here has been in development since 1998, arising from a need to analyze authorship attribution questions in forensic linguistics. This included exploring extant methods, developing novel methods of analysis and implementing methods in tools that embody those methods. It is convenient to conflate the notions of "methods" and "tools". Various aspects of the tools and analysis conducted using the tools have been published (Appel & Vogel, 2001; Van Gijsel & Vogel, 2003; O'Brien & Vogel, 2003; Vogel, 2007b; Healey, Vogel, & Eshghi, 2007; Vogel, 2007a; Vogel & Brisset, 2007; Vogel & Lynch, 2008; Frontini, Lynch, & Vogel, 2008; Vogel, Lynch, & Janssen, 2008). Student projects (e.g. McCombe, 2002) have been based on implementing pieces of a larger suite of tools, or testing those tools (Mencke, 2004; Medori, 2005; Gilmartin, 2006; Mencke, 2007). The particular toolset described here is a subset of the larger set of tools for stylistic analysis, a subset tuned to addressing a smaller set of questions. While aspects of the larger suite of methods have been published as just indicated, this component has not before been published.[1] The general form of the question is: how does a family of probed texts rank in relation to a reference corpus? The more specific form is: can sections of translated corpora be ranked according to overall resemblance with positive examples of linguistic style provided by a reference corpus? The evaluative term, "effectively", means in this case that the items at the top of the rank would be likely judged by human assessors as acceptable, and those at the bottom of the rank, most likely to benefit from human intervention, if any items require intervention at all. Different choices of reference corpus may be expected to yield correspondingly different rankings of items under consideration. Therefore, the choice of reference corpus should be related to the criteria of acceptability that would be deployed by any human assessor. However, target reference corpora may be available even where it is not possible to articulate clear assessment criteria.

This question can be made more specific in any number of directions. Firstly, the features analyzed may ultimately be about semantic content or they may address syntactic properties of texts. Either of those dimensions can be seen as falling within stylistic considerations. To the extent that one has faith in the representative nature of the reference corpus, whether homogeneous or not (Kilgarriff, 1997, 2001; Kilgarriff & Salkie, 1996; Vogel, 2007a; Vogel et al., 2008), one can have faith that the method described here provides a statistically oriented method of validating conformity between the probe text and reference tests – a statistical grammar and style checker. Importantly, there need not be a codified description of the reference style. Rather, the reference style is

[1] It is the subject of a patent application pending with the European Patent Office: EPO 11169673.8-1527 (2011) "Data processing system and method for assessing quality of a translation".

deemed implicit in reference corpora independently classified as conforming to the target style. The relevant aspects of style are made tangible through the sort of tokenization[2] used as the basis for $n-$gram distribution comparisons that yield ranks of relative stylistic conformity.[3]

The purpose of this article is to describe the details of the method. First, we review recent literature that has addressed discovery of translation effects. We then describe the details of our method, at the level of algorithmic description. Finally, we situate our method by contrast with prior literature.

2 Translation Effects

Translation effects that emerge in comparison of bodies of translated and original texts have been the subject of recent research in computational and corpus linguistics. This section expands on a recent overview provided by Lynch and Vogel (2012).

Olohan (2001) identifies patterns in *optional* usage in comparable English corpora.[4] For example, use of complementizer *that*[5] discriminates translations and original texts; translations contain a greater relative frequency of the complementizer construction. The method deployed by Olohan depends on selective expert hypotheses about which features discriminate texts of translated English from those originally composed in English, with t-tests used to identify those features which differ with statistical significance.

Guthrie, Guthrie, Allison, and Wilks (2007) evaluated their general method of ranked feature differences, among other evaluations, on the problem of assessing whether translations of L1 Chinese newspaper texts in L2 English could be identified in a set of L1 English news texts (35K words of Chinese translated to English and 50K words of English L1). The linguistic features were what we consider to be document-level features (i.e. percentages of words in major grammatical categories, ratios of frequencies between grammatical categories, most frequent POS trigrams and bigrams, etc). Feature vectors represent each text and its relative complement, with separate vectors for the percentages and ratios and the ranked frequency features. Derived vectors record a score based on the Spearman rank correlation coefficient between the text and its complement, for each of the sorts of frequency list. Texts are compared by calculating the average differences between feature vectors and adjusting with the derived scores from the ranked frequency list differences. In each configuration of the

[2]Tokenization is the process of individuating texts into countable components: letters, words, lemmas, parts of speech, etc. We count sequences of variable length: sequences of individual tokens are unigrams, of length two are bigrams, etc. and for any length n are n-grams.

[3]For Part of Speech item labels, we have used the IMS TreeTagger (Schmid, 1994). Tag analysis is not described further in this document.

[4]Comparable corpora have internal homogeneity of style and genre, and in this context contain proportional amounts of translated and original texts (see Kilgarriff (2001) on comparability assessment).

[5]*He said* ***that*** *he was ill* vs. *he said he was ill* vs. *the illness* ***that*** *killed him was swift*: the first contains a complementizer-that and the last, a relativizer-that.

evaluation, one translation was presented without annotation along with 50 L1 English texts, with texts separated as 1000 word samples. Translated text appeared in the top three ranked positions, corresponding to greatest anomaly, in 93% of experiments, and in the top ten positions in 100%. Lynch and Vogel (2012) explored comparable feature sets, attempting to guess the source language of each text, rather than giving each text a rank in its evidence of being a translation, in a more refined version of the translation guessing problem.

Baroni and Bernardini (2006) explore whether machine learning (ML) methods may discover translated texts more effectively than humans. For a corpus of translated and original articles from the Italian current affairs publication *Limes*, they found a high degree ($\geq$85%) of classification accuracy between the two categories, identifying features such as clitic pronouns and adverbial forms as distinguishing features between the translated and original sections of the corpus. Only one of ten humans in an evaluation exercise outperformed the ML system on all measures.

Seeking source language cues in the Europarl corpus van Halteren (2008) attempted automated detection of translations, obtaining high accuracy in L1 detection($\geq$ 90%) across translations and original texts in multiple European languages, using features such as n-grams of words and POS tags, without document-level features. Diagnostic n-grams included *framework conditions* in the English corpus translated from German, and the n-gram *certain number*, which occurred with greater relative frequency in translations from French and Spanish than in translations from the German, Italian and Dutch texts. More recent work by Ilisei, Inkpen, Corpas Pastor, and Mitkov (2010) on stylistics of translations in Spanish technical and medical translations motivates features at the document level, for example, "proportion of simple sentences, complex sentences and sentences without any finite verb" (Ilisei et al., 2010, p. 506).

Detection of L1 of a non-native speaker writing in a foreign language is a similar task to that of discovering translation effects, although of course professional translators generally translate into their native language. Wong and Dras (2009, 2011) use sentence parses and n-gram features[6] to detect syntactic idiosyncrasies in non-native speaker text, reporting 80% classification accuracy for seven different L1 types[7] using sentence parses and ˜70% accuracy using n-gram features only on a corpus of learner essays. In this case the corpus was highly comparable, consisting entirely of learner essays in English.

Kochmar (2011) adopts a similar approach to Wong and Dras (2011) in the task of L1 identification of non-native speaker English text, focusing on a number of two-class classification problems, including broader categories such as Romance languages (French, Italian, Catalan, Spanish, Portuguese) vs. Germanic languages (German, Swiss German, Dutch, Swedish and Danish) and more finely grained classifications including Spanish vs. Catalan, for example. Features used include word n-grams, POS n-grams and character n-grams, together with more complex syntactic features such as phrase structure rules and

[6]character n-grams, POS n-grams, function word frequencies.

[7]Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese

frequencies of different error types. Kochmar (2011) also obtains 84% classification accuracy for the Germanic vs. Romance task using a combination of character unigrams, bigrams and trigrams, POS unigrams, bigrams and trigrams and word unigrams as features and does not perform any multiclass classification experiments in her study, unlike our experiments (Lynch & Vogel, 2012) which attempt to classify four different L1s.

Brooke and Hirst (2012) develop an alternative method for the task of L1 classification from non-native English text, they obtain word-for-word translations of word trigrams from a large blog corpus of Chinese, French, Japanese and Spanish text and use these features and also features derived from these (e.g. bigrams and POS n-grams) as training data for classification of an author's native-language in corpora such as the International Corpus of Learner English and other similar collections of text. They report results above the baseline of 25% (48% using word bigrams on the ICLE test corpus) however conclude that the results are not accurate enough to advocate using their method as the sole metric for L1 classification.

Lynch and Vogel (2012) study translation effects in literary texts, attempting to guess the source language of texts which may or may not have actually been translated. Where the classification of the source language as other than English is correct, there is also a correct guess that the text is translated. The feature individuation is hybrid, both lexical sequences and document level features (averages and frequency ratios coputed over the document; see Table 1 and Table 2) are included.[8] The experiments reported yield 80% accuracy. As a matter of principle, if one can automatically identify the source language of a text, therefore guessing whether it is a translation, then one can use this ability to assess the quality of linguistic items that are meant to seem as if originally composed in the translation target language rather than to be noticeable as translations. Exploration of this possibility across text genres constitutes a large part of our ongoing research program.

Table 1: Document-level statistics used by Lynch and Vogel (2012)

Feature	**Description**
Avgsent	Average sentence length
Avgwordlength	Average word length
Automated Readability Index	Readability metric
Coleman Liau Index	Readability metric

[8]The Automatic Readability Index (ARI) is calculated from a regression equation with coefficients applied to average word length and average sentence length, and correlates non-linearly with US grade level on the K-12 educational progression model (Smith & Senter, 1967). The Coleman-Liau Index (CLI) is also based on a regression equation, with coefficients applied to letters per hunderd words and sentences per hundred words correlating linearly with a 16-level educational grade progression model (Coleman & Liau, 1975). Both are similar in spirit, essentially assessing average word-length and average sentence length.

Table 2: Document-level ratios used by Lynch and Vogel (2012)

Feature	Ratio Description
Typetoken	word types : total word tokens
Numratio	numerals : total word tokens
Fverbratio	finite verbs : total word tokens
Prepratio	prepositions : total word tokens
Conjratio	conjunctions : total word tokens
Infoload	open-class words : total word tokens
dmarkratio	discourse markers : total word tokens
Nounratio	nouns : total word tokens
Grammlex	open-class words : closed-class words
simplecomplex	simple sentences : complex sentences
Pnounratio	pronouns : total word tokens
lexrichness	lemmas : total word tokens
simplecomplex	simple sentences : complex sentences
simpletotal	simple sentences : total sentences
complextotal	complex sentences : total sentences

We next describe our method for integrating comparisons of texts with reference corpora to construct rankings relative to two parameters: comparison metric and reference corpus. As a matter of expository convenience, we describe the method with a fixed comparison metric (averaged χ^2 over word bigrams) noting that any comparison metric could be substituted. That is, one might use POS trigrams and cosine similarity, or document level metrics, and so on.

3 Description

3.1 System Architecture and Data Flow

The picture provided in Figure 1 demonstrates the steps in the analytical process. The hexagons denote processes that require implementation and are in some cases (hexagon-1) composites of further processes that merit further discussion (see §3.4.2). The rectangular boxes mark data files either presumed as inputs (e.g. the corpus, or the corpus index) or generated as important outputs (e.g. the frequency distribution index of the corpus according to the tokenization and sequence length adopted, the file by file similarity measures, the file by file similarity ranks, the aggregated rank of probed items across the reference categories). While we use the term "file", we mean this term to be neutral between whole or part of a document: one might use the analysis to rank sentences within a document, thereby assigning each to a file, or one might rank entire documents that comprise a translated corpus. In the figure, the "X3" is meant to indicate that trigrams of some sort of tokenization has been applied (e.g. X = "w" signals word tokenization; X = "w!" indicates word tokenization with

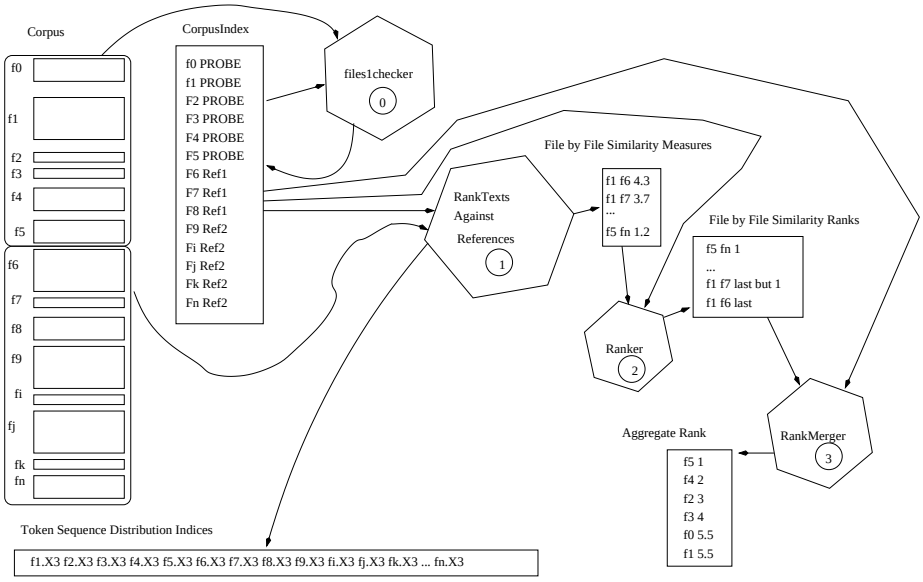

Figure 1: System Architecture and Data Flow

punctuation included as tokens; etc.). Other auxiliary information is included in the output of some of the analysis, and implementations should record these for future use; however, they do not merit further discussion here.

The diagram in Figure 1 omits aspects of the tool suite that are not relevant to the sort of question intended to be addressed with this subset of the tools. For example, a larger process takes into account the internal homogeneity of each category of text being considered and as such, it requires consideration not just of PROBE items with respect to each reference item, but also each reference item with respect to each other reference item.

3.2 Input Types and Data Types

3.2.1 Inputs

Corpora to be analyzed using these methods as implemented are presumed to be presented in raw textual format. If one is ranking sentences in a document then they are to be individuated as separate files, with an index maintained of each. As such, the scale of corpora to be ranked using these methods is variable—a corpus of documents can be ranked against other documents, or sentences within a document may be ranked.

The index of documents must contain the name of each file to be considered and an indication of whether it is a PROBE text or a reference text. A reference corpus may contain arbitrarily many subcategories. The ranking is conducted in three steps, and these steps construct outputs that become inputs to later parts

of the analysis. Certain input parameters influence behavior; for example one may signal that punctuation should be taken into account as distinct tokens.

3.2.2 Outputs

The primary output of interest is the aggregate ranking of each of the input probed items against the totality of the reference corpus. It is noted that the reference corpus may be comprised of sub-categories, and each of the input files is considered with respect to each of these categories in producing the aggregate ranking. In any case, any PROBE file is compared against all reference files, and ultimately it is necessary to construct an aggregated rank for each PROBE file on the basis of all its reference set comparisons.

It makes sense to construct aggregate rankings on the basis of preliminary rankings rather than on raw similarity scores. This is because the rankings do not preserve distance between two items, but the distance recorded by the similarity scores depends on the nature of each subcategory of reference items.

The similarity scores that we have explored most are derived from the χ^2 statistic. The details of this computation are spelled out below. However, any similarity metric between two items may be substituted in place of this particular computation. An output that complements similarity scores is indication of significant dissimilarity. For example, one might wish to think of documents as vectors, with an index for each token-sequence that might occur, given the analysis chosen, and then examine the cosine between vector representations of two items. One might wish to substitute other more holistic metrics of raw similarity between items as well.

3.3 Processing Assumptions

It is useful to verify that the index of files in the probe and in the reference are adequately populated with respect to tokenization in terms of words, and sequences of them. The main input parameters are the index label and the value that counts as a zero. That is, if one is processing using n-grams, then any file with n-1 tokens effectively has zero content. A secondary parameter allows automatic adjustment of the index to construct a new index eliminating items with effectively zero content. In addition to the essential data inputs, it is useful to consider whether tokenization should take case into account, or whether it should include punctuation marks as tokens.

3.4 Key Processes

3.4.1 Driving computations

In the first instance, it is necessary to assess frequency distributions with respect to the token sequence length of interest across the entire corpus of items being considered.

Given an index of items which indicates for each item whether it is a PROBE item or, if not, what reference category it belongs to (and presuming that the

integrity of this index has been checked so that less error checking is necessary during the processing of the items); and a further indexing of those items which records the total distribution of token sequences contained in each item: compute the average similarity score for each PROBE item in comparison with each reference item, recording these similarity scores (optionally, where similarities cross a significance threshold, record those anomalous component comparisons).

3.4.2 Average similarity score

This document details the use of average χ^2 ratio as the metric of similarity between two items. It is important enough to reiterate that the method could substitute an alternative method of computing similarities between items. Whichever similarity score is used, the computation as depicted in the system architecture and data-flow diagram of Figure 1 is located in the hexagon with the label, "Rank Texts against References" (hexagon-1).

The computation of a similarity score between two items using the χ^2 ratio involves considering the total distribution of token sequences that occur in both items. For each token sequence that occurs in either item, one computes the χ^2 score, and the similarity score for the two items is the average of these individual token sequence comparisons. A single token sequence comparison (e.g. word unigram) considers the observed frequency of the token sequence in the one item and its observed frequency in the other item, in relation to the expected frequency of the token sequence in each of those items. The expected frequency of the token sequence is determined by the total number of token sequences that comprise the item. Thus, one is essentially considering a series of two by two contingency tables, one for each token sequence that is instantiated by the comparison of two items. See Table 3, for example: τ represents the token sequence of focus for the table, and $\overline{\tau}$ represents the occurrences of each token sequence that is not the token sequence τ. A comparable table of expected values is derived from the observed values.

Observations of Token Sequence	Item 0	Item 1	Total Across Items
τ	a	b	a + b
$\overline{\tau}$	c	d	c + d
	a + c	b + d	n = a + b + c + d

Table 3: Table of observations for a single token sequence τ

Expected Instances of Token Sequence	Item 0	Item 1
τ	(a+b)*((a+c)/n)	(a+b)*((b+d)/n)

Table 4: Table of expected values for a single token sequence τ

If one imagines the first row of Table 3 as defining a function $o(\tau)$, then $o(\tau, 0)$ and $o(\tau, 1)$ pick out the observed values for τ in the first and second items, respectively, and Table 4 provides the method of computing the expected values on the basis of the observed values and in relation to the total size of the two items being compared.

$$\chi^2 = \frac{(o(\tau,0) - e(\tau,0))^2}{e(\tau,0)} + \frac{(o(\tau,1) - e(\tau,1))^2}{e(\tau,1)} \quad (1)$$

The method which we have explored most carries a cumulative sum of the χ^2 values for each token sequence inspected between two items, and then divides by the total number of distinct values for τ (N). Note that this value, minus one (N-1), is the equivalent of the degrees of freedom for the overall contingency table of observations of each τ, and "chi by degrees of freedom" refers to a method rather like this one, but with the divisor set to be N-1 rather than N (Oakes, 1998, p.28).

This accumulation of scores for the individual τ provides an aggregate similarity measure for the item. However, using the assumptions of the χ^2 test from inferential statistics, one can also comment along the way on whether the distribution of τ in two items being compared is significantly different—indications of distinctive token sequences, and in aggregation, distinctive items, are derivable for other similarity metrics one might use, as well.

Setting a critical value for χ^2 according to a probability of making an error of judgement to the effect that the two items sampled are not from the same population when in fact they are (i.e. the probability of being wrong in concluding that the items are significantly different with respect to some τ), if the number of observations is at least 5 in both cells,[9] then it is appropriate to signal that an anomalous token sequence within the comparison of items has been identified.

The process involves comparing each of the items in the PROBE category (P) with each of the reference items (R), and thus involves $O(P * R)$ item-level comparisons. Other sorts of processes that are useful to compose using the same similarity score, in assessing the homogeneity of each category and sub-category being analyzed, for example, require $O((P * R)^2)$ item-level comparisons. Thus, there are advantages to using the methods described here for text analysis problems that permit the efficient processing provided.

3.4.3 Rank

Note that ranking involves a standard treatment for assigning ranks to ties—if there are ties between ranks i and j ($j > i$), then assign to all such comparisons the rank $(\frac{(i+j)}{k})$ where k is the number of tied comparisons. This preserves the rank-sum property, namely that the sum of the ranks of the items in a ranked list of n items should be equal to $\frac{n(n+1)}{2}$.

[9] Some work with this constraint on expected values rather than on observed values.

Given an index of items which indicates for each item whether it is a PROBE item or, if not, what reference category it belongs to; and further given a sorted set of raw similarity scores that emerge from item by item comparisons: for each item by item comparison rank comparisons *in relation to all other pairwise comparisons*. This ranking is insensitive to reference sub-categories inasmuch as they are not treated separately in this step. Each comparison involves one PROBE item and one reference item, and each PROBE item has a score with respect to each reference item. This step adds rank information in satisfaction of the rank-sum property just described to the raw similarity scores obtained for each item by item comparison, according to the similarity metric and tokenization sequence (as described in § 3.4.2).

In Figure 2 the output of this process is the file-by file rankings in the rectangular box between hexagon-2 (for the ranker) and hexagon-3 (for rank-merging). Note that this figure is an expansion of a subset of Figure 1. Figure 2 omits mention of the individual items in both PROBE category and each reference sub-category; it also omits the tokenized index of the individual items output from the preparatory phases of the processes that computes file-by-file similarity scores (hexagon-1). Only one pass through these files and the indices is necessary using this method. The focus is, rather, on the ranker (hexagon-2) and the rank-merger (hexagon-3).

The output of this ranking is reduced further in two directions. Firstly, one wants to know the rank of each PROBE item not just with respect to each reference item as much as with respect to each reference category. Secondly, one typically wants to abstract over this and assess each PROBE item in terms of its ranking across each of the reference categories.

3.4.4 Aggregate Rank Merge

Given an index of items which indicates for each item whether it is a PROBE item or, if not, what reference category it belongs to;[10] and further given a *ranked* set of raw similarity scores that emerge from item by item comparisons:[11] for each reference category, rank each comparison in relation to the other comparisons for that reference category;[12] and construct an aggregate rank of PROBE items across the reference categories.[13]

It would be natural to consider taking the input sorted and ranked raw similarity scores that derive from comparisons of items, and aggregating those similarity scores directly. However, it is not clear whether the mathematical operations presupposed in the direct aggregation of raw scores retains face va-

[10] In Figure 2, this is the *CorpusIndex* box at the left of the diagram.

[11] This is the output of the ranker.

[12] In Figure 2, this is the rectangle with the tab-label "A". The dotted line from the output of the ranker to items within this rectangle is meant to illustrate the process: the rank of an item with respect to a sub-category depends on the sum of the item's rank on the basis of each item in the sub-category; so, the most similar two files were f5 and fn, the latter of which is a member of f2, and this contributes the value 1 to rank sum for f5 with Ref2, and the comparison of f5 with fj contributes the value rank-k, and so on.

[13] This is the final ranked list of outputs—in Figure 2, the rectangle with the tab-label "C".

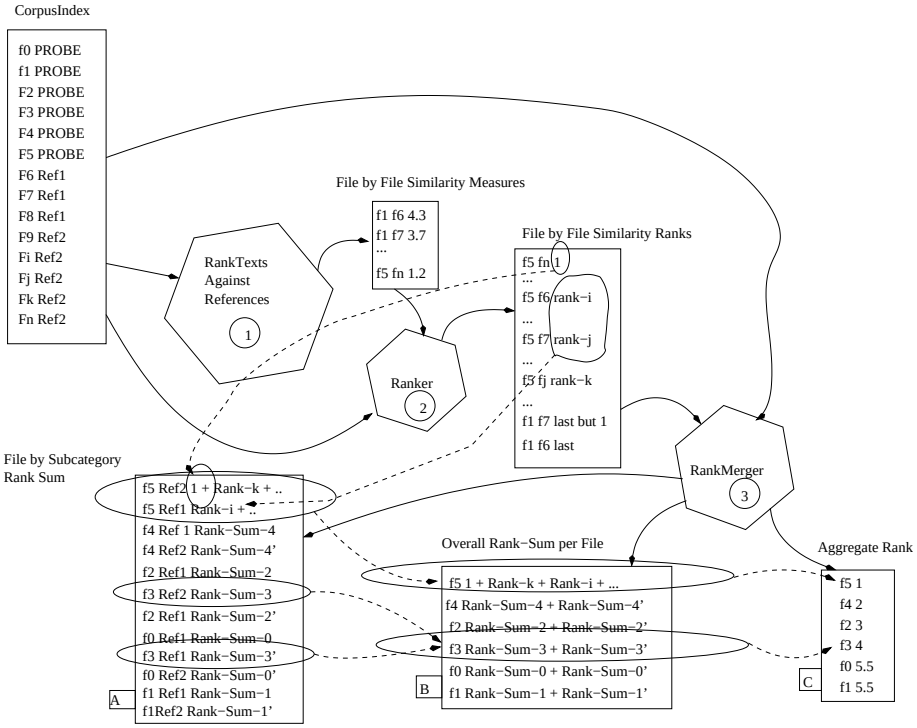

Figure 2: Ranking items and Merging Ranks

lidity (this is a separate matter from whether as an approximation strategy, it works). One may produce a distance-preserving aggregation of similarity of items with other items into similarity of items with subcategories and then with the overall reference category. The raw scores record the distance between comparisons, and this is exactly what is lost in abstracting raw scores into ranks. An issue is that the relative distances obtained by similarity scores change scales between items because the pairs of items may contain different numbers of tokens sequences. It is natural enough to average the similarity scores for token sequence distributions within an item by item comparison, but the motivation for averaging across item by item comparison is not clearly valid for all metrics of similarity.[14] However, research into non-parametric statistics has provided compelling tools for reasoning with statistics based on rank orderings, and the work described here is in the spirit of that research.

While the distance between points in comparisons across a sub-category or across the entire reference corpus is not to an obvious normalization that preserves metric properties, the rank ordering of comparisons, which preserves relative ordering of points but not distance between points, is suitable for ag-

[14]Note that this is an open question: the sentence footnoted here does not say, "the motivation for averaging ... is clearly not valid."

gregation across multiple comparisons. Therefore, for each PROBE item, and for each reference sub-category, the sum of the ranks obtained for the PROBE item paired with each of the items in the reference sub-category is computed (this is depicted in the rectangle with the tab-label "A", in Figure 2). It is reasonable to think of this as the PROBE item's rank sum within each sub-category. Thus, one has the information which can be sorted and handed on to rank each PROBE item with respect to the other PROBE items in the context of each reference sub-category. Then, for each PROBE item, the sum of its rank-sums across reference sub-categories is computed (this is depicted in the rectangle with the tab-label "B", in Figure 2; notice that the rank of f5 with respect to Ref1 and Ref2 are added together, as shown by the values enclosed in an ellipse in "A" and connected by a dashed arrow to "B"). This models the PROBE item's rank sum across each sub-category. These sum-of-rank scores are sorted then handed on to be ranked in satisfaction of the rank-sum property (this is depicted in the rectangle with the tab-label "C", in Figure 2). This is the effective output of the method.

Detailing a straightforward variant of this method clarifies what the method amounts to: it would have been possible, further, to additionally construct the average rank position for each item by simply dividing the sum-of-ranks by the total number of reference sub-categories, and handing this information on for ranking to satisfy the rank-sum property again. The base method discussed here constructs scores simply through sum-of-rank information rather than dividing that by the total number in the relativization that would yield an average.

4 The structure of the analytical process

In deploying the methods described here, it is necessary to amass a body of suitable reference texts for the task at hand. In the case of translation quality control, the reference texts would consist of documents in the target language deemed to be of an acceptable standard for comparison. It is not necessary that there be a codified "house style". Rather, the house style may be implicit in the reference set of positive examples that are deemed to follow the style; a complement set of negative, non-conforming, examples is not needed. It is also necessary to decide how to treat the item that is being subjected to a quality-control analysis. The item may be a class of documents or a decomposition of a single document into constituent parts (paragraphs or sentences provide natural decompositions). The treatment consists of separating the item into units at the level of granularity required and indexing those units as items to PROBE. On a model in which one decomposes a larger text into the sentences it contains, the method instantiates a sentence-level statistical style and grammar checker.

Suppose, for example (see Figure 3), that a new translation of *The Odyssey* is offered. In evaluating it, one might decompose it into N segments corresponding more or less to sentences. One might compare this new translation with the translations of the Greek epics by Robert Graves and Alexander Pope. It does not matter particularly that *The Iliad* is a distinct poem, and it would not

matter enormously if Pope's translation were actually in prose. The method as constructed supports the identification of the j of the N segments below a cut-off point in the ranking which are least similar to the overall corpus of translations of Greek epics into English, for example. Thresholds can be set depending on whether one wants to favor precision or recall in the identification of items that are candidates for further inspection.[15] It is a separate matter that one could identify which sub-categories of the reference material are less like the others, using the outputs of the proposed method of analysis. This can be quite helpful in assessing the reference material on hand, actually. In the example depicted in Figure 3, the initial data-verification step is not shown. The figure intends to illustrate the process of assembling a reference corpus and decomposing a document to be probed into files for individual items within the document (the upper left of the diagram), through the analytical steps described in this document, towards the ranking of the items probed and determination of a subset of those items below a cut-off point which merit closer inspection for conformity with respect to the reference corpus.

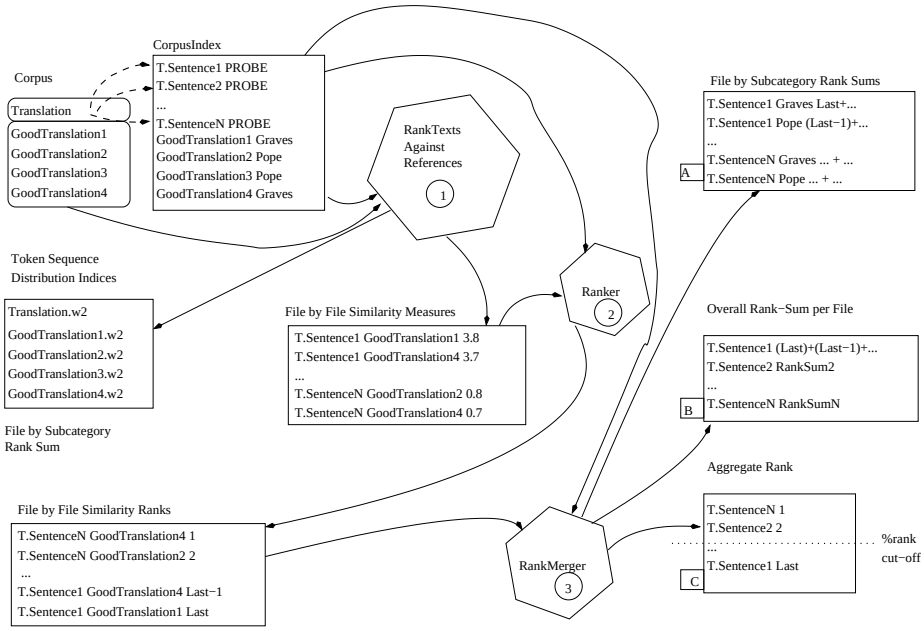

Figure 3: Hypothetical Example – evaluating a new translation of *The Odyssey*

[15] In empirical tests using translations into Russian of corporate material, we have identified the bottom 40% of the items ranked as a good place to look for candidate items in need of further examination of quality as translations.

5 Discussion

This section elaborates on how this recent research is situated with respect to the literature. The main claim here is that it is possible to achieve with automated means, using a computer, a process that would otherwise be impossible to achieve, namely, comprehensive evaluation of the entirety of a vast set of translated documents for conformity with respect to a reference corpus. The resulting ranking of elements allows a more precise specification of items requiring manual inspection for style and grammaticality than has hitherto been feasible. The method is independent of the choice of human language in which the corpora are presented. The approach is robust in being applicable across problem categories (not merely translation quality assurance), with effectiveness, prior to tuning to any tuning for particular problems. In contrast, methods that require greater levels of expertly annotated training data require more tuning to particular problem categories before their efficacy is evident.

This question can be made more specific in any number of directions. Firstly, note that in making the judgements depend on a set of items to be probed against a reference corpus, one may use the method to analyze items in any language with orthography; thus, settling on a corpus does not entail settling on a language, since corpora may be heterogeneous with respect to language composition. Secondly, the features analyzed may ultimately be about semantic content or they may address syntactic properties of texts. Either of those dimensions can be seen as falling within stylistic considerations. To the extent that one has faith in the representative nature of the reference corpus, whether homogeneous or not (Kilgarriff, 1997, 2001; Kilgarriff & Salkie, 1996; Vogel, 2007a; Vogel et al., 2008), one can have faith that the method described here provides a statistical grammar and style checker. Importantly, there need not be a codified description of the reference style, no explicit reference grammar. The relevant aspects of style are implicit in the sort of tokenization to which the ranking is put and the choice of reference texts. References texts are plentiful.

Some past work on linguistic quality analysis have used statistical methods in various locations within the problem. For example, some work targets identifying what proportions of errors are made by individuals as a function of whether they are speakers of standard or non-standard dialects (Hagen, Johannessen, & Lane, 2001; Johannessen, Hagen, & Lane, 2002; Nerbonne & Wiersma, 2006). The method of Nerbonne and Wiersma (2006), for example, fixes on part of speech assignments to words used, and deploys computationally intensive permutation tests to assess significant differences in distributions between native speakers of Australian English and Finnish emigrant Australian English. Some work uses corpus driven analysis in order to locate gaps in hand crafted precision grammars (which may be thought of as "rule-based", in contrast to stochastic grammars) (Baldwin, Beavers, Bender, Flickinger, Kim, & Oepen, 2005). Primarily rule-based systems may contain components which are driven by statistical information—for example, a grammar checker by Knutsson (2001) has a component which guesses part of speech information to assign to words on the basis of statistical information, but a rule oriented component for

constructing linguistic generalizations. Grammar checkers have also been used for assessing translation outputs, at least in the case of machine translation: Stymne and Ahrenberg (2010) use the grammar checker of Knutsson (2001) for this purpose, conducting error analysis of machine translation output. Linguistic error analysis has been an explicit goal of some research, irrespective of the linguistic source (see e.g. Foster and Vogel (2004)). Stochastic grammars have also been developed for the specific purpose of grammar checking; however, work like that of (Alam, UzZaman, & Khan, 2006) ultimately presupposes a binary view of grammaticality, and records as grammatical evidently any sentence which has greater than zero probability according to the language model. Similarly, (Henrich & Reuter, 2009) exploit the fact also exploited in the work here that a purely statistically driven system may be language independent; however, their notion of grammaticality is still binary in that token sequences which do not occur in their equivalent of the reference corpus results in ungrammaticality. On the other hand, much work in linguistic theory presupposes (Ross, 2004, (orig. 1973)) or explores a more graded notion of grammaticality (Vogel & Cooper, 1994; Frank, King, Kuhn, & Maxwell, 1998; Aarts, Denison, Keizer, & Popova, 2004; Crocker & Keller, 2006; Fanselow, Féry, Vogel, & Schlesewsky, 2006) (including gradience in syntactic category/part of speech assignment (Aarts, 2007)).

With reference to this background, the method for producing an evaluation of grammar and style of translated texts through a conformity-ranked presentation of probed items as described here is: automated and efficient; language independent; eschews a rule-based element of grammar or style analysis; open to token-sequence distribution analysis between probes and references for any number of levels of tokenization; capable of enumerating locations of significant differences between the reference corpus and probe items; enables evaluation of the entirety of the corpus to be probed. In this article we have not reported detailed evaluations of the method proposed. However, we have separately studied the method through evaluation experiments in work that is unpublished. In these preliminary studies, the method successfully ranked as among the least similar to the reference corpus translated items that were independently selected for human post-editing. These studies lead us to new research questions, for example, in establishing levels of lexical overlap between probe items and reference corpora as a parameter for interpretation of rank similarity results. The success in these preliminary studies also led us to incorporate the process as a component stage in a larger machine translation process: within a larger collaboration we used the architecture to evaluate texts at the input end of translation in an attempt to improve data driven machine translation where limited training resources exist, using our methods to select training data that is most similar to source text categories to be translated. Essentially, the task involved machine translation of online forum texts for which directly relevant translation training data was not available. However, other data sources, like Europarl, contain parallel translations of some transcripts which are closer to online forum data than are software manuals for which translation memories do exist. Our method of selecting training items from the "most similar" end of the similarity

ranking, in comparison with a control method of random sampling (without replacement) of an equal number of training items, yielded significantly better machine translation evaluation metrics. This material is in preparation for peer review. In the context of linguistic quality estimation, we have evaluated weakly supervised methods more generally,[16] and have verified that methods like this which do not require prior detailed classification of individual training sentences perform well, particularly where the trade-off costs associated with training are high (Moreau & Vogel, 2012, 2013).[17]

6 Conclusion

This article provides a review of automated approaches to detecting translation effects alongside the details of a specific method for analyzing stylistic differences between translation and target language reference corpora. A feature of the work is that there need not be an explicit style guide for the target corpora, nor is it necessary to compile a reference corpus of negative examples. Stylistic conformity is based on intersubstitutable similarity metrics between probe items and reference corpora. Multiple rankings may be aggregated. Items which rank low in aggregate similarity may be regarded as in need of additional human intervention. The benefit of the approach is that it makes it possible to evaluate the entirety of translation projects automatically, highlighting those items within the project that are most distant in stylistic similarity from the reference corpus. Those items would be the ones, if any,[18] that might benefit from further human intervention. We have also described our ongoing research program in improving automated assessment of translation quality against the family of baseline methods supplied by this research. One element of this program is determining the parameters of data sets that suggest which metrics are applicable (or inapplicable), in the same sense that parametric inferential statistics are applicable or not based on whether observed variables appear normally distributed. Another element is in the competing paradigm of labelling rather than ranking as we have expounded here: in our labelling approach, it is any item that is classifiable as a translation (or in stronger terms, any item for which we can automatically guess the source language on the basis of linguistic features alone) that merits human re-examination.

[16] The method we describe here is best characterized as weakly supervised.

[17] Our method relies on the availability of reference texts and demands no more than coarse grained categorization of the reference texts; in contrast, more strongly supervised learning methods depend on high quality annotations of items that serve for training. At present, oceans of reference texts are available via the Internet for typical uses of our methods, in comparison to ponds of annotation data, custom constructed in general, for strongly supervised approaches.

[18] All items may be of high quality, but relative to the reference corpus, the items are nonetheless ranked; thus, the bottom tiers of the rank may require no further attention.

References

Aarts, B. (2007). *Syntactic Gradience: The Nature of Grammatical Indeterminacy*. Oxford University Press.

Aarts, B., Denison, D., Keizer, E., & Popova, G. (Eds.). (2004). *Fuzzy Grammar: A Reader*. Oxford University Press.

Alam, M. J. A., UzZaman, N., & Khan, M. (2006). N-gram based statistical grammar checker for bangla and english. In *In Proc. of Ninth International Conference on Computer andInformation Technology (ICCIT 2006)*.

Appel, C. & Vogel, C. (2001). Investigating syntax priming in a e-mail tandem language learning environment. In Cameron, K. (Ed.), *C.A.L.L. — The Challenge of Change: Research and Practice*, pp. 177–84. Exeter: Elm Bank Publications.

Baldwin, T., Beavers, J., Bender, E., Flickinger, D., Kim, A., & Oepen, S. (2005). Beauty and the beast: what running a broad-coverage precision grammar over the bnc taught us about the grammar – and the corpus. In Kepsar, S. & Reis, M. (Eds.), *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, pp. 49–69. Mouton De Gruyter. Studies in Generative Grammar 85.

Baroni, M. & Bernardini, S. (2006). A new approach to the study of translationese: machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259.

Brooke, J. & Hirst, G. (2012). Measuring interlanguage: native language identification with l1-influence metrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* Istanbul, Turkey. European Language Resources Association (ELRA).

Coleman, M. & Liau, T. (1975). A computer readability formula designed for machine scoring.. *Journal of Applied Psychology*, *60*(2), 283.

Crocker, M. & Keller, F. (2006). Probabilistic grammars as models of gradience in language processing. In Fanselow, G., Féry, C., Vogel, R., & Schlesewsky, M. (Eds.), *Gradience in Grammar: Generative Perspectives*, pp. 227–245. Oxford University Press.

Fanselow, G., Féry, C., Vogel, R., & Schlesewsky, M. (Eds.). (2006). *Gradience in Grammar: Generative Perspectives*. Oxford University Press.

Foster, J. & Vogel, C. (2004). Parsing ill-formed text using an error grammar. *Artificial Intelligence Review*, *21*, 269–291.

Frank, A., King, T. H., Kuhn, J., & Maxwell, J. (Eds.). (1998). *Optimality Theory Style Constraint Ranking in Large-Scale LFG Grammars*. Stanford: CSLI Publications. `http://csli-publications.stanford.edu/LFG/3/lfg98-toc.html`.

Frontini, F., Lynch, G., & Vogel, C. (2008). Revisiting the donation of constantine. In Kibble, R. & Rauchas, S. (Eds.), *2008 Artificial Intelligence and Simulation of Behavior – Symposium: Style in Text*, pp. 1–9.

Gilmartin, E. (2006). An investigation of linguistic coordination in dialogue between native speakers of english and russian-speaking learners of english.. MPhil in Linguistics, Centre for Language and Communication Studies, Trinity College, University of Dublin.

Guthrie, D., Guthrie, L., Allison, B., & Wilks, Y. (2007). Unsupervised anomaly detection. In *IJCAI*, pp. 1624–1628.

Hagen, K., Johannessen, J. B., & Lane, P. (2001). Some problems related to the development of a grammar checker. In *NODALIDA'01: 2001 Nordic Conference in Computational Linguistics*.

Healey, P. G. T., Vogel, C., & Eshghi, A. (2007). Group dialects in an online community. In Arnstein, R. & Vieu, L. (Eds.), *DECALOG 2007, The 10th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 141–147. Università di Trento (Italy), May 30 – June 1, 2007.

Henrich, V. & Reuter, T. (2009). Lisgrammarchecker: language independent statistical grammar checking. Master's thesis, Hochschule Darmstadt and Reykjavík University.

Ilisei, I., Inkpen, D., Corpas Pastor, G., & Mitkov, R. (2010). Identification of translationese: a machine learning approach. *Computational Linguistics and Intelligent Text Processing*, 503–511.

Johannessen, J. B., Hagen, K., & Lane, P. (2002). The performance of a grammar checker with deviant language input. In *Proceedings of the 19th international conference on Computational linguistics - Volume 2*, COLING '02, pp. 1–8 Stroudsburg, PA, USA. Association for Computational Linguistics.

Kilgarriff, A. & Salkie, R. (1996). Corpus similarity and homogeneity via word frequency. In *Proceedings of Euralex 96*.

Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proc. 5th ACL SIGDAT Workshop on Very Large Corpora*, pp. 231–245.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, *6*(1), 97–133.

Knutsson, O. (2001). Automatisk språkgranskning av svensk text.. Licentiate Thesis, Royal Institute of Technology (KTH), Stockholm, Sweden.

Kochmar, E. (2011). Identification of a writers native language by error analysis. Master's thesis, University of Cambridge.

Lynch, G. & Vogel, C. (2012). Towards the automatic detection of the source language of a literary translation. In Kay, M. & Boitet, C. (Eds.), *24th International Conference on Computational Linguistics*, Vol. 1, pp. 775–784. COLING2012: Posters; Mumbai, India, 8-15 December 2012 (33% acceptance rate in total for the sum of posters and long papers).

McCombe, N. (2002). Methods of author identification. B.A. (Mod) CSLL Final Year Project, TCD.

Medori, J. (2005). Experiments testing the reliability and validity of author identification methods. Master's thesis, Computational Linguistics Group, Trinity College Dublin. Computational Linguistics Group.

Mencke, M. (2004). Experiments to validate scientifically reliable author identification techniques.. MPhil in Linguistics, Centre for Language and Communication Studies, Trinity College, University of Dublin.

Mencke, M. (2007). Benchmarking a text classification technique. Master's thesis, Computational Linguistics Group, Trinity College Dublin.

Moreau, E. & Vogel, C. (2012). Quality estimation: an experimental study using unsupervised similarity measures. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pp. 120–126. Association for Computational Linguistics.

Moreau, E. & Vogel, C. (2013). Weakly supervised approaches for quality estimation. *Machine Translation*. Accepted – to appear.

Nerbonne, J. & Wiersma, W. (2006). A measure of aggregate syntactic distance. In Nerbonne, J. & Hinrichs, E. (Eds.), *Linguistic Distances Workshop*, pp. 82–90. COLING/ACL.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.

O'Brien, C. & Vogel, C. (2003). Spam filters: bayes vs. chi-squared; letters vs. words. In et al., M. A. (Ed.), *Proceedings of the International Symposium on Information and Communication Technologies*, pp. 298–303.

Olohan, M. (2001). Spelling out the optionals in translation: a corpus study. *UCREL Technical Papers*, *13*, 423–432.

Ross, J. R. (2004). Nouniness. In Aarts, B., Denison, D., Keizer, E., & Popova, G. (Eds.), *Fuzzy Grammar: A Reader*, pp. 351–422. Oxford University Press. Originally in Osamu Fujimura (ed.) *Three Dimensions of Linguistic Research*. Tokyo: TEC Ltd. (1973):137-257.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pp. 44–49. Manchester, UK.

Smith, E. & Senter, R. (1967). Automated readability index.. Tech. rep. AMRL-TR-66-220, Aerospace Medical Research Laboratories (6570th).

Stymne, S. & Ahrenberg, L. (2010). Using a grammar checker for evaluation and postprocessing of statistical machine translation. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pp. 2175–2181.

Van Gijsel, S. & Vogel, C. (2003). Inducing a cline from corpora of political manifestos. In et al., M. A. (Ed.), *Proceedings of the International Symposium on Information and Communication Technologies*, pp. 304–310.

van Halteren, H. (2008). Source language markers in europarl translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 937–944. Coling 2008 Organizing Committee.

Vogel, C. (2007a). Corpus homogeneity and bernoulli schema. In *Mining Massive Data Sets for Security*, pp. 93–94. NATO Advanced Study Institute.

Vogel, C. (2007b). N-gram distributions in texts as proxy for textual fingerprints. In Esposito, A., Keller, E., Marinaro, M., & Bratanic, M. (Eds.), *The Fundamentals of Verbal and Non-Verbal Communication and the Biometrical Issue*, pp. 189 – 194. Amsterdam: IOS Press.

Vogel, C. & Brisset, S. (2007). Hearing voices in the poetry of brendan kennelly. *Belgian Journal of English Language & Literature*, 1–16.

Vogel, C. & Cooper, R. (1994). Robust chart parsing with mildly inconsistent feature structures. In AAAI Technical Report: AAAI 1994 Fall Symposium—*Knowledge Representation for Natural Language in Implemented Systems*. November 4-6, 1994, New Orleans, LA. A revised version appears in Andreas Schöter and Carl Vogel (eds.) *Nonclassical Feature Systems*. Edinburgh Working Papers in Cognitive Science, Volume 10. (1995) EUCCS-WP10. pp. 197-216.

Vogel, C. & Lynch, G. (2008). Computational stylometry: who's in a play. In Esposito, A., Bourbakis, N., Avouris, N., & Hatzilygeroudis, I. (Eds.), *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, Vol. LNCS-5042/2008, pp. 169–186. Berlin: Springer.

Vogel, C., Lynch, G., & Janssen, J. (2008). Universum inference and corpus homogeneity. In Bramer, M., Coenen, F., & Petridis, M. (Eds.), *AI-2008 Twenty-eighth SGAI International Conference on Artificial Intelligence*, pp. 367–372. Springer.

Wong, S. & Dras, M. (2009). Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*, p. 53. ALTA.

Wong, S. & Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1600–1610. Association for Computational Linguistics.