Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words

Oren Tsur
Institute of Computer Science
The Hebrew University
Jerusalem, Israel
oren@cs.huji.ac.il

Ari Rappoport
Institute of Computer Science
The Hebrew University
Jerusalem, Israel
arir@cs.huji.ac.il

Abstract

We apply modern statistical NLP techniques to study language transfer, a major issue in the theory of Second Language Acquisition (SLA). Using an SVM for the problem of native language classification, we show that a careful analysis of the effects of various features can lead to substantial scientific insights. In particular, we demonstrate that character bi-grams alone allow classification levels of about 66% for a 5-class task even when content and function word differences are accounted for. We hypothesize that the phonology of a native language has a strong effect on the word choice of people writing in a second language.

1 Introduction

While advances in NLP achieve improved results for NLP applications such as machine translation, question answering and document summarization, there are other fields of research that can benefit from the methods developed by the NLP community. Second Language Acquisition (SLA), a major area in Applied Linguistics and Cognitive Science, is one such field. In this paper we demonstrate how modern NLP tools can lead to a substantial insight in SLA theory. In particular, we address the major SLA issue of language transfer (interference), the effect of native language on second language learners. Using an SVM for the computational problem of native language classification, we study in detail the effects of various SVM features. Surprisingly, character bi-grams alone lead to a classification accuracy of about 66% in a 5-class task, even when accounting for differences in content and function words.

This result leads us to form a novel hypothesis regarding the role of language transfer in second language acquisition: that the choice of words people make when writing in a second language is strongly influenced by the phonology of their native language.

This is the first time that such a hypothesis has beed formulated. Moreover, this is the first statistical learning-related hypothesis in language transfer, and modern statistical NLP techniques are absolutely essential in supporting it. Clearly, such a claim must be further substantiated by additional psycholinguistic and computational experiments; nonetheless, we provide a strong starting point.

The next section provides some background. In Section 3 we describe our experimental setup and feature selection, and in Section 4 we detail an array of variations of experiments for ruling out some possible types of bias that might have affected the results. In Section 5 we locate our hypothesis within psycho-linguistic theory. We conclude by suggesting directions for future research.

2 Background

Our hypothesis is tested within an algorithm addressing the practical problem of determining the native language of an anonymous writer writing in a foreign language. The problem is applicable to different fields, such as language instructing, tailored error correction, security applications and psycholinguistic research.

As background, we start from the somewhat related problem of authorship attribution. The authorship attribution problem was addressed by linguists and other literary experts trying to pinpoint an anonymous author, such as that of The Federalist Papers (Holmes and Forsyth, 1995). Traditionally, authorship experts analyzed topics, stylistic idiosyncrasies and personal information about the possible candidates in order to determine an author. While authorship is usually addressed with deep human inspection of the texts in question, it has already been shown that automatic text analysis based on various stylistic features can identify the gender of an anonymous author with accuracy above 80% (Argamon et al., 2003; Koppel et al., 2003). Various papers (Dietrich et al., 2003; Koppel and Schler, 2003; Koppel et al., 2005a; Stamatatos et al., 2004) report relative success in machine based authorship attribution tasks for small sets of known candidates.

Native language detection is a harder problem than the authorship attribution problem, since we wish to characterize the writing style of a set of writers rather than the unique style of a single person. There are several works presenting non-native speech recognition and dialect analysis systems (Bouselmi et al., 2005, Bouselmi et al., 2006; Hansen et al. 2004). However, all those works are based on acoustic signals, not on written texts. Koppel et al. (2005a) report an accuracy of 80% in the task of determining a writer's native language. To the best of our knowledge, this is the only published work on automated classification of an author's native language (along with another version of the paper by the same authors (Koppel et al., 2005b)). Koppel et al. used a combination of features in their system (such as errors analysis and POS-error co-occurrences, as described in section 2.2), but surprisingly, it appears that a very nave set of features achieves a relatively high accuracy. The character bi-gram frequencies feature performs rather well, and definitely outperforms the intuitive contribution of frequent bigrams in this type of task.

3 Experimental Setting

3.1 The Corpus

The corpus that served for all of the experiments described in this paper is the International Corpus of Learner English (ICLE) (Granger et al. 2002), which was also the one used by Koppel et al. (2005a; 2005b). The corpus was compiled for the purpose of studying the English writing of non-native speakers. All contributors to the corpus are advanced English students and are roughly the same age. The corpus is combined from a number of sub-corpora, each containing one native language. The corpus was assembled in ten years of international collaboration between a number of universities and it contains more than 2 million words of writing by students from 19 different native language backgrounds. We followed Koppel et al. (2005a) and worked on 5 sub-corpora, each containing 238 randomly selected essays by native speakers of the following languages: Bulgarian, Czech, French, Russian and Spanish. Each of the texts in the corpus was written by a different author and is of length between 500 to 1,000 words. Each of the sub corpora contains about 180,000 (unique) types, for a total of 886,677 tokens.

Essays in the corpus are of two types: argumentative essays and literature examination papers. Descriptive, narrative or technical subjects were not included in the corpus. The literature examination essays were restricted to no more than 25% of each sub-corpus. Each contributor was requested to fill a learner profile that was used to fine-proof the corpus as needed.

In order to verify our results we used another control corpus containing the Dutch and Italian sub-corpora contained in the ICLE instead of the Bulgarian and French ones.

3.2 Document Representation

In the original experiment by Koppel et al. (2005a) each document was represented by a numerical vector of 1,035 dimensions. Each vector entry represented the frequency (relative to the document's length) of a given feature. The features were of 4 types:

- 400 function words
- 200 most frequent letter n-grams
- 250 rare POS bi-gram
- 185 error types

While the first three types of attributes are relatively straightforward, the fourth is more complex. It represents clusters of families of spelling errors as well as co-occurrences of errors and POS tags. Document representation is described in detail in (Koppel et al., 2005a; Koppel et al., 2005b).

A multi-class SVM (Witten and Frank, 2005) was employed for learning and evaluating the classification model. The experiment was run in a 10-fold cross validation manner in order to test the effectiveness of the model.

3.3 Previous Results

Koppel et al. (2005a) report that when all features types were used in tandem, an accuracy of 80.2% was achieved. In the discussion section they analyze the frequency of a few function words, error types, the co-occurrences of POS tags and errors, and the co-occurrences of POS tags and certain function words that seem to have significance in the support vectors learnt by the SVM.

The goal of their research was to obtain the best classification, therefore the results obtained by using only bi-grams of characters were not particularly noted, although, surprisingly, representing each document by only using the relative frequency of the top 200 characters bi-grams achieves an accuracy of about 66% . We believe that this surprising fact exposes some fundamental phenomenon of human language behavior. In the next section we describe a set of experiments designed to isolate the causes of this phenomenon.

4 Experimental Variations and Results

Intuitively, we do not expect the most frequent character n-grams to serve as good native language predictors, expecting that these will only reflect the most frequent English words (and characters sequences). Accordingly, without language transfer effects, a nave baseline classifier based on an n-gram model is expected to achieve about 20% of accuracy in a 5 native languages classification task. However, using classification based on the relative frequency of top 200 bi-grams achieves about 66% [1] in all experiments, substantially higher than the random baseline. These results are so surprising that they suggest that the characters bi-grams classification masks some other bias or noise in the corpus, or, conversely, that it mirrors other simple-to-explain phenomena such as shallow language transfer through the use of function words or content bias. The following sub-sections describe different variations of the experiment, ruling out the effect of these different types of bias.

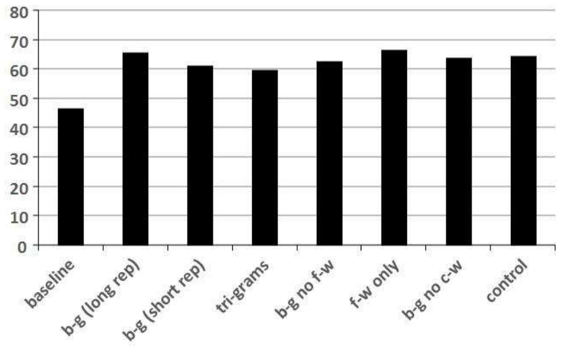

Figure 1: Classification accuracy of the different variations of document representation. b-g: bi-grams, f-w: function words, c-w: content words.

4.1 Unigram Baseline

We first implemented a nave baseline classifier. We represented each document by the normalized frequencies of the (de-capitalized) letters it contains [2]. These frequencies are simply a unigram model of the sub-corpora. Using the multi-class SVM (Witten and Frank, 2005) we obtained 46.78% accuracy. This accuracy is more than twice the random baseline accuracy. This result is in accorddance with our bi-grams results. Our discussion focuses on bi-grams rather than unigrams because the former's results are much higher and because bi-grams are much closer to the phonology of the language, approximating syllables rather than consonants and vowels (for alphabetic scripts, of course).

4.2 Bi-grams Based Classification

Choosing the 200 most frequent character bi-grams in the corpus, we used a vector of the same dimension. Each vector entry contained the normalized frequency of one of the bi-grams. Using a multi-

[1] Koppel et. al. did not report those results explicitly. However, they can be roughly estimated from their graph.

[2] White spaces were considered a letter. However, sequences of white spaces and tabs were collapsed to a single white space. All the experiments that make use of character frequencies were performed twice, including and excluding punctuation marks. Results for both experiments are similar, therefore all the numbers reported in this paper are based on letters and punctuation marks.

	Bulgarian	**Czech**	**French**	**Russian**	**Spanish**
dr	**170**	183	n/a	195	n/a
am	**117**	135	142	140	152
m_	121	120	133	**119**	139
iv	**104**	138	144	148	148
_y	**161**	181	196	183	166
la	122	123	122	142	**105**

Table 1: some of the separating bi-grams that were found in the feature selection process. '_' indicates a white space. The numbers are the frequency ranking of the bi-grams in each sub-corpus (e.g., there are 103 bi-grams more frequent than 'iv' in the Bulgarian corpus). n/a indicates that this bi-gram is not one of the 200 most frequent bi-grams of the sub-corpus.

class SVM in a 10-fold cross validation manner we achieved 65.60% accuracy with standard deviation of 3.99.

The bi-grams features in the 200 dimensional vector are the 200 most frequent bi-grams in the whole corpus, regardless of their frequency in each sub-corpus.

A more sophisticated feature selection could reduce the dimension of the representation vector without detracting from the results. Careful feature selection can also give a better intuition regarding the support vectors. We performed feature selection in the following manner: we chose the top 200 bi-grams of each sub-corpus, getting 245 unique bi-grams in total. We then chose all the bi-grams that were ranked significantly higher or significantly lower in one language than in at least one other language, assuming that those bi-grams have strong separating power. With the threshold of significance set to 20 we obtained 84 separating bi-grams. Table 1 shows some of the separating bi-grams thus found. For example, 'la' is a good separator between Russian and Spanish (its rank in the Spanish corpus is much higher than that in the Russian corpus), but not between other pairs.

Using only those 84 bigrams we obtained classification accuracy of 61.38%, a drop of only 4% comparing to the results achieved with the 200 dimensional vectors. These results show that pumping the dimension of the representation vector using additional bi-grams contribute a marginal improvement while it does not introduce substantial noise.

4.3 Using Tri-gram Frequencies as Features

Repeating the same experiment with the top 200 tri-grams, we obtained an accuracy of 59.67%, which is 40% higher than the expected baseline and 15% higher than the uni-grams baseline. These results show that the texts in our corpus can be classified by only using nave n-gram models, while the optimal n of the n-gram is a different question that might be addressed in a different work (and might be language-dependent).

4.4 Function Words Based Classification

Function words are words that have a little lexical meaning but instead serve to express grammatical relations within a sentence or specify the attitude of the speaker (function words should not be confused with stopwords, although the lists of most frequent function words and the stopword list share a large subset). We used the same list of 460 function words used by Koppel et al. (2005a). A partial list includes: {*a, afterward, although, because, cannot, do, enter, eventually, fifteenth, hither, hath, hence, lastly, occasionally, presumable, round, said, seldom, undoubtedly, was*}.

In this variation of the experiment, we represented each document only by the relative frequency of the function words it contained. Using the same experimental setup as before, we achieved an accuracy of 66.7%. These results are less surprising than the results obtained by the character n-grams vectors, since we do expect native speakers of a certain language to use, misuse or ignore certain function words as a result from language transfer mechanisms (Odlin, 1989). For example, it is well known that native speakers of Russian tend to omit English articles.

4.5 Function Words Bias

The previous results suggest that the n-gram based classification is simply the result of the different uses of function words by speakers of different native languages. In order to rule out the effect of the function words on the bi-gram-based classification, we removed all function words from the corpus, recalculated the bi-gram frequencies and ran the experiment once again, this time achieving an accuracy of 62.92% in the 10-fold cross validation test.

These results, obtained on the function words-free corpus, clearly show that n-gram based classification is not a mere artifact masking the use of function words.

4.6 Content Bias

Bi-gram frequencies could also reflect content bias rather than language use. By content bias we mean that the subject matter of the documents in the different sub-corpora could exhibit internal sub-corpus uniformity and external sub-corpus disparity. In order to rule this out, we employed a variation on the Term Frequency - Inverted Document Frequency (*tf-idf*) content analysis metric.

The *tf-idf* measure is a statistical measure that is used in information retrieval tasks to evaluate how important a word/term is to a document in a collection or corpus (Salton and Buckley, 1988). Given a collection of documents D, the *tf-idf* weight of term t in a document $d \epsilon D$ is computed as follows:

$$tfidf_t = f_{t,d} \times log\frac{|D|}{f_{t,D}}$$

where $f_{t,d}$ is the frequency of term t in document d, and $f_{t,D}$ is the number of documents in which t appears. Therefore, the weight of term $t \epsilon d$ is maximal if it is a common term in d while the number of documents it appears in is relatively low.

We used the *tf-idf* weights in the information retrieval sense in order to discover the dominant content words of each sub-corpus. We treated each sub-corpus (set of documents by writers who share a native language) as a single document and calculated the *tf-idf* of each word. In our experiments we have used a common though less standard version of the *tf-idf* which is more suitable to our task – omitting the *log* operator at the *idf* part of the *tf-idf* function. In order to determine whether there is a content bias or not, we set a dominance threshold, and removed all words such that the difference between their *tf-idf* score in two different sub-corpora is higher than the dominance threshold. Given a threshold *t, the dominance* $D_{w,t}$*, of a token* w is given by:

$$D_{w,t} = max(tfidf_{w,i} - tfidf_{w,j})$$

where $tfidf_{w,k}$ is the *tf-idf* score of token w in sub-corpus k. Changing the threshold in 0.0005 intervals, we removed from 1 to 340 unique content words (between 1,545 and 84,725 word tokens in total). However, the classification accuracy was essentially the same (see Figure 2), with a slight drop of only 2% after removing 51 content words (by using a threshold of 0.0015).

word	Bulgarian	Czech	French	Russian	Spanish
europe	0	0.0003	0.0027	0.0002	0.0002
european	0	0.0003	0.003	0.0001	0.0005
imagination	0.0043	0.002	0.0008	0.001	0.0008
television	0	0.0036	0.0019	0.0031	0.0003
women	0.0004	0.0017	0.0012	0.0055	0.0026

Table 2: *tf-idf* score of some of the most dominant words

Subcorpus	content words	function words	unique stems
Bulgarian	1543	94685	11325
Czech	2784	110782	12834
French	2059	67016	9474
Russian	2730	112410	12338
Spanish	2985	108052	12627
Total	12101	492945	36474

Table 3: number of dominant content words (with a threshold of 0.0025) and function words that were removed from each sub-corpus. The unique stems column indicates the number of unique stems (types) that remained after removal of *c-w* and *f-w*.

We calculated the *tf-idf* weights after stop-words removal and stemming (using a Porter stemmer (Porter, 1980)), trying to pinpoint dominant stems. The results were similar to the word's *tf-idf* and no significantly dominant stem was found in either of the sub-corpora.

A drop of only 3% in accuracy was noticed after removing both dominant content words and function words. These results show that if a content bias exists in the corpus it has only a minor effect on the SVM classification, and that the n-grams based clas-

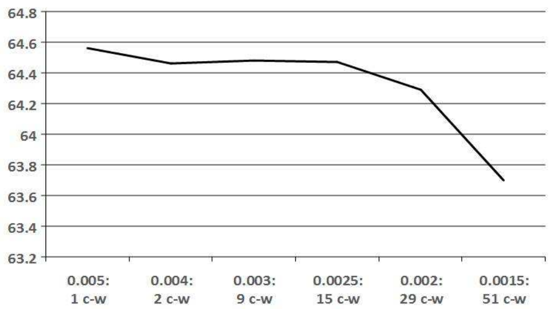

Figure 2: Classification accuracy as a function of the threshold (removed content words).

Threshold	0.004 2 *c-w*	0.003 9 *c-w*	0.0025 15 *c-w*	0.0015 51 *c-w*	0.001 113 *c-w*
Bulgarian	77	908	1543	3955	7426
Czech	306	1829	2784	5139	8588
French	665	1829	2059	3603	6205
Russian	781	1886	2730	6302	9918
Spanish	389	1418	2985	6548	10521
Total	2218	7970	12101	25547	42658

Table 4: number of occurrences of content words that were removed from each sub-corpus for some of the thresholds. The numbers in the top row indicate the threshold and the number of unique content words that were found with this threshold.

sification is not an artifact of a content bias.

We ran the same experiment five more times, each time on 4-sub-corpora instead of 5, removing one (different) language each time. The results in all 5 4-class experiments were essentially the same, and similar to those of the 5 language task (beyond the fact that the random baseline for the former is 25% rather than 20%).

4.7 Suffix Bias

Bias might also be attributed to the use of suffixes. There are numerous types of English suffixes, roughly speaking, they may be divided to morphemic and grammatical (inflectional). It is reasonable to expect that just like a use of function words, use or misuse of certain suffixes might occur due to language transfer and first language interference. A frequent use of a certain suffix or avoidance of the use of a certain suffix may influence the bi-grams statistics and thus the bi-grams classification may be only an artifact of the suffixes usage.

Checking the use of the 50 most productive suffixes taken from a standard list (e.g. *ing, ed, less, able, most, en*) we have found that only a small number of suffixes are not equally used by speakers of all 5 languages. Most notable are the differences in the use of *ing* between native French speakers and native Czech speakers and the differences of use of *less* between Bulgarian and Spanish speakers (Table 5). However, no real bias can be attributed to the use of any of the suffixes because their relative aggregate effect on the values in the support vector entries is very small.

suffix	Bulgarian	Czech	French	Russian	Spanish
ing	872	719	932	903	759
less	47	36	39	45	32

Table 5: count of two of the suffixes whose frequency of use differs the most between sub-corpora.

4.8 Control Corpus

Finally, we have also ran the experiment on a different corpus trading the French and the Spanish sub-corpora with the Dutch and Italian ones, introducing a new Roman language and a new Germanic language to the corpus. We obtained 64.66% accuracy, essentially the same as in the original 5-language setting.

The corpus was compiled from works of advanced English students of the same level who write essays of approximately the same length, on a set of randomly and roughly equally distributed topics. We expected that those students will use roughly the same n-grams distribution, however, the results described above suggest that there exists some mechanism that dictates the authors' choice of words. In the next section we present a computational psycholinguistic framework that might explain our results.

5 Statistical Learning by Infants and Language Transfer in SLA

5.1 Statistical Learning by Infants

Psychologists, linguists, and cognitive science researchers try to understand the process of language learning by infants. Many models for language learning and cognitive language modeling were suggested (Clark, 2003).

Infants learn their first language by a combination of speech streams, vocal cues and body gestures. Infants as young as 8 months old have a limited grasp of their native tongue as they react to familiar words. In that age they already understand the meaning of single words, they learn to spot these words in a speech stream, and very soon they learn to combine different words into new sentential units. Parental speech stream analysis shows that it is impossible to separate between words by identifying sequences of silence between words. (Saffran, 2001). Recent studies of infant language learning are in favor of the statistical framework (Saffran, 2001; Saffran et al., 1996). Saffran (2002) examined 8 month-old to one year-old infants who were stimulated by speech sequences. The infants showed a significant discrimination between word and non-word stimuli. In a different experimental setup infants showed a significant discrimination between frequent syllable n-grams and non frequent syllable n-grams, heard as part of a gibberish speech sequence generated by a computer according to various statistical language models. In a third experimental setup infants showed a significant discrimination in favor of English-like gibberish speech sequences upon non-English-like gibberish speech sequences. These findings along with the established finding (Jusczyk, 1997) that infants prefer the sound of their native tongue suggest that humans learn basic language units in a statistical manner and that they store some statistical parameters pertaining to these units. We should note that there are people who doubt these conclusions (Yang, 2004).

5.2 Language Transfer in SLA

The role of the first language in second language acquisition is under a continuous debate (Ellis, 1999). Language Transfer (LT) between L1 and L2 is the process in which a language learner of L2 whose native language is L1, infers from L1 when using L2 (actually, when building his/her inter-language model). This inference might appear helpful when L2 is relatively close to L1, but most likely it interferes with the learning process due to over- and under-generalization or other problems. Although there is clear evidence that language learners use constructs of their first language when learning a foreign language (James, 1980; Odlin, 1989), it is not clear that the majority of learner errors can be attributed to the L1 transfer and to L1 interference (Ellis, 1999).

Language transfer is usually associated with L1 interference (negative transfer) that causes syntactical errors. While some people believe that syntax is learned and used statistically (at least partially) (Bybee, 2006), this is far from being an accepted view.

5.3 Syllable Sound Transfer Hypothesis

We hypothesize that there are language transfer effects influenced by L1 syllable sounds and manifested by the words that people choose to use when writing in a second language. (We say 'writing' because we have only experimented with written texts; a more general hypothesis covering speaking and writing can be formulated as well.)

Furthermore, since the acquisition and representation of syllable sound is strongly influenced by statistical considerations (Section 5.1), we speculate that the general language transfer phenomenon might be related to frequency. This does not follow from our findings, of course, but is an exciting direction to investigate.

We note that there is one obvious and well-known lexical transfer effect - using cognates (words that have similar form (sound) and meaning in two different languages). However, the languages we used in our experiments contain radically differing amounts of cognates of English words (just consider French vs. Bulgarian, for example), while the classification results were about the same for all 5 languages. Hence, cognates might play a role, but it is not the major factor explaining our findings.

We note that the hypothesis put forward in the present paper is the first that attributes a language transfer phenomenon to a cognitive representation whose statistical nature has been seriously substantiated.

6 Conclusion

In this paper we have demonstrated how modern NLP techniques can aid other fields, here the important field of Second Language Acquisition (SLA). Our analysis of the features useful for a multi-class SVM in the task of native language classification has resulted in the formulation of a hypothesis of large potential significance in the theory of language transfer in SLA. We hypothesize language transfer effects at the level of sound sequences (specifically, syllables, reflected as two-letter sequences), manifested by the words that people choose when writing in a second language. In other words, the use of L2 words is strongly influenced by the frequency of L1 syllables.

As noted above, further experiments (psychological and computational) must be conducted for validating our hypothesis. In particular, construction of a wide-scale learners' corpus (possibly including additional native languages) with tight control over content bias is essential in order to reach stronger conclusions. However, the effort involved might be beyond the capability of a single research group.

Additional future work should address syllable sequences vs. the orthographic sequences that were used in this work. If our hypothesis is correct, then using transcripts of spoken language corpora written in a phonetic script should produce even stronger results (our results are affected by the fact that writing systems rarely show a 1-1 correspondence with how words are at the phonological level). Our eventual goal is creating a unified model of statistical transfer mechanisms.

References

Argamon S., Koppel M. and Shimoni A. 2003. *Gender, Genre, and Writing Style in Formal Written Texts*. Text 23(3).

Bouselmi G., Fohr D., Illina, I., and Haton J.P. 2005. *Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model*. Eurospeach/Interspeach. 2005.

Bouselmi G., Fohr D., Illina I., and Haton J.P. 2006. *Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration and Graphemic Constraints*. IEEE International Conference on Acoustics, Speech and Signal Processing, 2006.

Bybee J. 2006. *Frequency of use and the organization of language*. Oxford University Press.

Clark, E. 2003. *First Language Acquisition*. Cambridge University Press.

Diederich J., Kindermann J., Leopold E. and Paass G. 2004. *Authorship Attribution with Support Vector Machines*. Applied Intelligence, 109-123.

Ellis R. 1999. *Understanding Second Language Acquisition*. Oxford University Press.

Granger S., Dagneaux E. and Meunier F. 2002. *International Corpus of Learner English*. Presses universitaires de Louvain.

Hansen J. H., Yapanel U., Huang, R. and Ikeno A. 2004. *Dialect Analysis and Modeling for Automatic Classification*. Interspeach-2004/ICSLP-2004: International Conference Spoken Language Processing. Jeju Island, South Korea.

Holmes D. and Forsyth R. 1995. *The Federalist revisited: New directions in authorship attribution*. Literary and Linguistic Computing, pp. 111-127.

James C. E. 1980. *Contrastive Analysis*. New York: Longman.

Jusczyk P. W. 1997. *The Discovery of Spoken Language*. Cambridge, MA. MIT Press.

Koppel M. 2006. *New Methods for Attribution of Rabbinic Literature*. Hebrew Linguistics: A Journal for Hebrew Descriptive, Computational and Applied Linguistics..

Koppel M. and Schler J. 2003. *Exploiting Stylistic Idiosyncrasies for Authorship Attribution*. in Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis. Acapulco, Mexico.

Koppel M., Schler J. and Zigdon K. 2005. *Determining an Author's Native Language by Mining a Text for Errors*. Proceedings of KDD 2005. Chicago IL.

Koppel M., Schler J. and Zigdon K. 2005. *Automatically Determining an Anonymous Author's Native Language*. In Intelligence and Security Informatics (pp. 209-217). Berlin/ Heidelberg: Springer..

Odlin T. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press.

Porter F. M. 1980. *An algorithm for suffix stripping*. Program, Vol. 14, No. 3, 130-137.

Saffran J. R. 2001. *Words in a sea of sounds: The output of statistical learning*. Cognition, 81, 149-169.

Saffran J. R. 2002. *Constraints on statistical language learning*. Journal of Memory and Language, 47, 172-196.

Saffran J. R., Aslin R. N. and Newport E. N. 1996. *Statistical learning by 8-month old infants*. Science, issue 5294, 1926-1928.

Salton G. and Buckley C. 1988. *Term-weighing approaches in automatic text retrieval*. Information Processing and Management, 24(5): 513-523.

Stamatatos E,. Fakotakis N. and Kokkinakis G. 2004. *Computer-Based Authorship Attribution Without Lexical Measures*. Computers and Humanities, 193-214.

Witten I. H. and Frank E. 2005. *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

Yang C. 2004. *Universal grammar, statistics, or both?*. Trends in Cognitive Science 8(10):451-456, 2004.