LEXICALIZING COMPUTATIONAL STYLISTICS
For Language Learner Feedback

Julian BROOKE and Graeme HIRST
University of Toronto, Department of Computer Science,
10 King's College Road, Toronto, Ontario, Canada M5S 3G4
{jbrooke,gh}@cs.toronto.edu

1 Background

1.1 Computational Stylistics

Computational stylistics refers informally to a collection of tasks within computational linguistics that deal with the style—as opposed to the semantic content—of natural language. The most famous of these tasks is perhaps authorship attribution (Stamatatos et al., 2001), which uses statistical variations in word choice to select the most likely from a fixed set of potential authors. Though applicable to a number of different applications, this framework has been applied specifically to literary analysis, grouping authors by their style (Luyckx et al., 2006). A broader definition of style brings it closer to the definition of register or genre (Biber and Conrad, 2009), which has also received some attention in the context of text classification (Kessler et al., 1997). In educational and literacy contexts, the readability of a text is an important stylistic feature, and the automatic detection of grade level, for instance, has been addressed in recent work (Petersen and Osendorf, 2009). The field of text generation was among the first in computational linguistics to implement sensitivity to style (Hovy, 1990), which has continued into the modern, data-driven era (Paiva and Evans, 2005).

Though the tasks mentioned above are quite diverse, there are some common themes across computational work in style. First, there is a overwhelming focus on part-of-speech, function words, and textual statistics (e.g. word and sentence length) as the primary indicators of style. These types of features are attractive, of course, because they are exactly those features which would not, presumably, be overly influenced by the content of the text. They also appear in abundance, and so are easily leveraged by machine learning classification systems. In fact, these features have become so associated with style that they are often referred to as stylistic features. Lexical features, though common in other classification work, do not usually play a major role. A notable exception is Aragamon et al. (2007), which applies lexical features to a number of stylistic text classification tasks, but that work relies entirely on examples from linguistics textbooks, which we would argue cannot provide sufficient coverage. Our research agenda seeks to use automated methods to build large-coverage stylistic lexicons that represent various stylistic dimensions and that can be useful for various tasks. We focus, however, on the task of providing language learner feedback, because learners require much more precise, human-interpretable stylistic information than is necessary for the other tasks mentioned here.

1.2 Stylistic Feedback for Language Learners

There is a large body of work associated with computational tools for helping language learners, many of them focused on grammatical error correction, either using rule-based methods (Heift and Schulze, 2007) or modern statistical approaches (Leacock et al., 2010). Automated essay scoring systems (Shermis and Burstein, 2003) provide a starting point for more holistic, multi-aspect feedback, and there are ongoing studies showing that students do generally benefit from automated feedback (Grimes and Warschauer, 2010). These methods have been criticized, however, for failing to provide *construct validity* (Chung and Baker, 2003), that is, for relying on proxy features that do not necessarily reflect human judgments of quality.

To show the limitations of stylistics as it has been understood in this context, we briefly review the stylistic module included in the *Criterion* student feedback system, which is based on the ETS *e-rater* automated essay scorer (Attali and Burstein, 2006). The features that this module uses are detailed in Quinlan et al. (2009). They include: extreme sentence length; sentences beginning with a conjunction; the use of passive voice; the use of any of a small set of inappropriate words (expletives); and repetition, as determined by a statistical module (Burstein and Wolska, 2003). Though these features may serve to identify a small set of novice author errors, the range is extremely limited, and is focused on expert writer pet peeves, which may not reflect actual language use. For example, Quinlan et al. (2009) report that although the use of passive voice was originally intended as a negative feature, it was actually positively correlated with a higher human essay score. Besides, the simplistic good/bad style dichotomy offered by these kinds of feedback systems can frustrate students who wish to develop their own styles (Chen and Cheng, 2008). Instead, we would like to explore the space of possible stylistic dimensions that could be quantified using automated techniques.

1.3 Stylistic Dimensions

Although we reject the simple good/bad style approach to stylistic feedback, there is nevertheless much that can be learned about the possible dimensions of style by synthesizing the advice provided by prescriptive approaches to style (Fowler and Fowler, 1906; Strunk and White, 1979; Kane, 1983; Williams, 1990). Aspects of style that appear consistently in this genre include: clarity, simplicity, formality, concreteness, objectivity, naturalness, and many others. Though often vague, prescriptive linguistics nonetheless reflects commonsense understanding of stylistic effects; using the terminology of prescriptivism might be a good way to make feedback easily understood to human users.

In more empirically-grounded approaches to the definition of style (or register), the aspects of style often correspond to the objective facts of the communicative situation. For instance, Crystal and Davy (1969) define a set of ‘dimensions of situational constraint’, which include basic background information of the participants and discourse-specific information such as medium, topic, genre, and status (which predicts formality). The *Field*, *Tenor*, and *Mode* breakdown of systemic functional linguistics (Halliday, 1994) is a similar formulation. The model of Leckie-Tarry (1995) is explicitly based on the notion of various clines, correlated with each other via the main cline of register (oral/literate). Finally, dimensions of style can be derived via a bottom-up approach; classic work by Biber (1998) uses factor analysis to identify key dimensions of variation in the Brown corpus, including informational versus involved, situation-dependent versus context-independent, and narrative versus non-narrative. The technique that Biber uses is closely related to our method for deriving lexical formality.

2 Building a Lexicon of Formality

Among the dimensions uncovered by the review of relevant work, we chose *formality* as our first stylistic dimension to lexicalize. Formality, which is related to the notion of interpersonal distance as well as social status, is explicitly or implicitly referred to in much stylistic work, yet there are few computational approaches that deal with it explicitly. For more details on the method discussed here, see Brooke et al. (2010).

2.1 Latent Semantic Space

We apply a technique, *latent semantic analysis* or *LSA* (Landauer and Dumais, 1997) that has been used previously to build lexicons of (positive and negative) sentiment (Turney and Littman, 2003). First, we view each word as a vector of ones and zeros, corresponding to its appearance or absence in a large collection of documents; our documents here consist of a large corpus of blogs that have been pulled from the internet (Burton et al., 2009). Since there are millions of documents in the corpus, this vector is millions of bits long. However, a technique from linear algebra (*singular value decomposition*) allows us to reduce the number of dimensions to any fixed number *k*, and these new word vectors are guaranteed to be the best possible representation of the original variation (across words) that is possible in those *k* dimensions. This new *k*-dimensional space is referred to as latent semantic space, because it is able to generalize over the full document space, identifying the latent factors which can often correspond (roughly) to semantic (i.e. topic) variation. Stylistic variation appears in this latent semantic space as well, however; in fact, our work suggests that formality is a fundamental variation in a mixed-register corpus, since we identify formality best when we use low values of *k*. Note, though, that we do not find there to be a single latent variable (dimension) corresponding exclusively to formality. So instead, we create one.

2.2 Calculating Formality Scores

Once we have word vectors of *k*-dimensional latent semantic space, we create a formality metric by measuring the distance between each word in our vocabulary and a small set of "seed" words. Seed words are prototypical examples of the variation we are interested in. Our informal seeds were taken primarily from an online slang dictionary (e.g. *wuss*, *grubby*) and also include some contractions and interjections (e.g. *cuz*, *yikes*). The formal seeds were selected from a list of discourse markers (e.g. *moreover*, *hence*) and adverbs from a sentiment lexicon (e.g. *preposterously*, *inscrutably*); these sources were chosen to avoid words with overt topic, and to ensure that there was some balance of sentiment across formal and informal seed sets, so as to avoid creating a subjective/objective lexicon instead. We believe, however, we were only partially successful in that regard, which motivates our approach in Section 4. A standard distance metric is the cosine of the angle between the two vectors, and this is what we use here. A 2-dimensional example is shown in Figure 1.

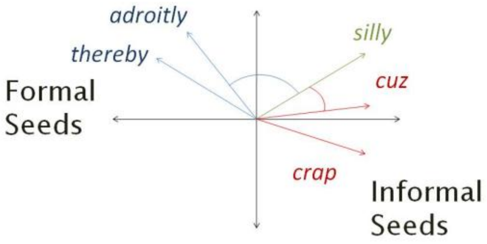

Figure 1. Two-dimensional angles from seeds of the word *silly*

For each word we average the distances and then normalize all the words to the range –1 to 1, providing a formality score.

2.3 Evaluation

We evaluated our technique using a set of word pairs derived from *Choose the Right Word* (Hayakawa, 1994), a manual for assisting writers in word choice among synonyms; all these pairs were either explicitly or implicitly compared for formality in the book. For various reasons, our pairs are biased toward the formal end of the spectrum; although there are some informal comparisons, e.g. *bellyache/whine*, *wisecrack/joke*, more typical pairs include *determine /ascertain* and *hefty/ponderous*. Our LSA-based dictionary built using a large web corpus correctly reflects the relative formality of these pairs over 80% of the time, and this can be boosted to over 85% if other information is included (i.e. word length and relative frequency in written and spoken corpora). We have also tested this method in other languages (Chinese and Finnish), and it could be easily adapted to other dimensions of style. If we have a list of multiword expressions (lexical bundles), these can also be assigned a formality score using the same technique. Once formality is quantified, we can easily identify words and expressions that deviate from the overall formality of the text.

3 Identifying Non-Nativeness Using L1 Texts

3.1 Rationale

Since our interest is in assisting language learners, it would be useful to point out to these learners their use of expressions which show influence from their native languages, i.e. language transfer (Odlin, 1989). Since the semantics of words are often given priority when learning mappings from one language to another, the style is often "lost in translation". However, we cannot apply the methods of the previous section because (L1-specific) non-nativeness is not a regular stylistic variation that would appear in a standard (mostly native) corpus, particularly when all the different possible native languages are considered (besides, it is unclear what our seed terms would be). Using manually collected non-native corpora such as the *International Corpus of Learner English* (*ICLE*) (Granger et al., 2009) is unlikely to provide enough data to find lexical indicators of non-native usage. Instead, we leverage the lexical information that can be derived from L1 texts (which are plentiful for most L1s), providing a quantification of L1-influence for words and word combinations. A full system would both identify these L1 influences and offer stylistically appropriate alternatives. We note also that a system which knows about patterns in the L1 has the potential to build trust with the user, who may not otherwise be able recognize that the system understands their original stylistic intent.

3.2 Method

The blog corpus whose English portion was used for formality lexicon acquisition also contains a significant quantity of non-English texts; We take 100 million words for each of four languages, viz. French, Spanish, Chinese, and Japanese. In order to go from L1 to L2 (English), we need a bilingual lexicon for each language; fortunately, such lexicons can be accessed on the web. Our software steps through each pair of contiguous words in the L1 texts (after some rearrangement to mimic English word order), and uses the bilingual lexicon to create an L2 equivalent based on a direct, literal translation. For example, the Chinese phrase 吃药 literally means "eat medicine", however the appropriate English form is "take medicine"; using our method, we can count appearances of 吃药 in Chinese texts, and thus learn that "eat medicine" is a sign of Chinese-influenced English. The level of Chinese-influence for a phrase is calculated using to the ratio of counts in Chinese as compared to all other L1s. Another advantage of this method is that we can remember that 吃药 was the source of "eat medicine", which also us to reconstruct the original intent of the author, and provide a better (more stylistically appropriate) expression.

3.3 Evaluation

We have tested the efficacy of these influence metrics by using them to identify L1s. Most work in native language identification (Koppel et al., 2005) involves supervised classification using machine learning algorithms, which are effective but have major disadvantages; they take advantage of variation unrelated to the problem of interest, and tend to do poorly in new corpora. Averaging our L1-influence metrics across all the word pairs in L2 texts, we have found that our metrics can predict L1 in the ICLE reasonably well (nearly 50% on the four-way task), better than using machine learning trained with a separate small corpus of L2 texts. More generally, lexical features seem key to the task, outperforming standard 'stylistic' features.

4 Stylistic Topic Models

In this section we briefly outline some future work. Though we could use LSA to derive other stylistic dimensions (as we did for formality in Section 2), the correlation among kinds of stylistic variation means that it is difficult to isolate particular dimensions independently. Instead, we would like to identify all dimensions in a single model. One promising approach is probabilistic topic modelling (Blei et al., 2003). Like LSA, topic models can be viewed as a dimensionality reduction technique that identifies latent variables (usually topics); unlike LSA, however, there is a great deal of flexibility in the underlying probabilistic models. For instance, latent variables can be correlated (Blei and Lafferty, 2007), which fits well into a stylistic framework (for instance, subjectivity and informality are likely to be correlated). There is a distinct challenge associated with this, however: how do we filter out the influence of topic to focus on stylistic variation, while still preserving our lexical focus?

5 Conclusion

In this paper, we have discussed an ongoing project to provide language learners automatically generated information about the stylistic connotations of lexical features, focusing on the derivation of two stylistic dimensions, formality and non-nativeness.

References

Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., and Levitan, S. (2007) Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 7:91–109.

Attali, Y. and Burstein, J. (2006) Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Biber, D. (1988) *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. and Conrad, S. (2009) *Register, Genre, and Style*. Cambridge University Press.

Blei, D.M. and Lafferty, J.D. (2007) Correlated topic models. *Annals of Applied Statistics*, 1(1):17–35.

Blei, D.M., Ng, A.Y., Jordan, J.I., and Lafferty, J.D. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research,* 3:993–1022.

Brooke, J., Wang, T., and Hirst, G. (2010) Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10).*

Burstein, J. and Wolska, M. (2003) Toward evaluation of writing style: finding overly repetitive word use in student essays. In *Proceedings of the 10th Conference of European Chapter of the Association for Computational Linguistics (EACL '03).*

Burton, K., Java, A., and Soboroff, I. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009),* San Jose, CA.

Chen, C.E. and Cheng, W.E. (2008) Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology,* 12(2):94–112.

Chung, G.K. and Baker, E.L. (2003) Issues in the reliability and validity of automated scoring of constructed responses. In Shermis, M.D. and Burstein, J., editors. *Automated Essay Scoring: A Cross Disciplinary Approach*. Lawrence Erlbaum Associates.

Crystal, D. and Davy, D. (1969) *Investigating English Style*. Indiana University Press.

Fowler, H. W. and Fowler, F. G. (1906) *The King's English*. Clarendon Press, 2nd edition.

Granger, S., Dagneaux, E., Meunier, F., and Paquot M. (2009) *International Corpus of Learner English (Version 2)*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Grimes, D. and Warschauer, M. (2010) Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment,* 8(6).

Hayakawa, S.I., editor (1994) *Choose the Right Word*. HarperCollins Publishers, 2nd edition, revised by Eugene Ehrlich.

Halliday, M.A.K. (1994) *Introduction to Functional Grammar*. Edward Arnold, 2nd edition.

Heift, T. and Schulze, M. (2007) *Errors and Intelligence in Computer-Assisted Language Learning*. Routledge, New York.

Hovy, E.H. (1990) Pragmatics and natural language generation. *Artificial Intelligence*, 43:153–197.

Kane, T.S. (1983) *The Oxford Guide to Writing*. Oxford University Press.

Kessler, B., Nunberg, G., and Schütze, H. (1997) Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97)*, pages 32–38.

Koppel, M., Schler, J., and Zigdon, K. (2005) Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, pages 624–628.

Landauer, T.K. and Dumais, S. (1997) A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Leacock, C., Chodorow, M., Gamon, M., and Tetreault, J. (2010) *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool.

Leckie-Tarry, H. (1995) *Language and Context: A Functional Linguistic Theory of Register*. Pinter.

Luyckx, K., Daelemans, W., and Vanhoutte, E. (2006) Stylogenetics: Clustering-based stylistic analysis of literary corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*.

Odlin, T. (1989) *Language Transfer*. Cambridge University Press.

Paiva, D.S. and Evans, R. (2005) Empirically-based control of natural language generation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pages 58–65.

Petersen, S.E. and Ostendorf, M. (2009) A machine learning approach to reading level assessment. *Computer Speech and Language*, 23(1):89–106.

Quinlan, T., Higgins, D., and Wolff, S. (2009) Evaluating the construct-coverage of the e-rater scoring engines. Technical report, Educational Testing Service.

Shermis, M.D. and Burstein, J., editors. (2003) *Automated Essay Scoring: A Cross Disciplinary Approach*. Lawrence Erlbaum Associates.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2001) Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35:193–214.

Strunk, W. and White E.B. (1979) *The Elements of Style*. Pearson Education, 3rd edition.

Turney, P. and Littman, M. (2003) Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Williams, J. M. (1990) *Style: Towards Clarity and Grace*. University of Chicago Press.