Wang, Tong and Hirst, Graeme
Department of Computer Science, University of Toronto
{tong, gh}@cs.toronto.edu

Associating Difficulty in Near-Synonymy Choice with Types of Nuance Using Core Vocabulary

There are arguably infinitely many dimensions along which members of a cluster of near-synonyms can differ (Cruse, 1986), and the nuances that differentiate near-synonyms along a dimension are often subtle and difficult even for native speakers. The diversity of near-synonym variation types has motivated the categorization of these dimensions of these variations. DiMarco, Hirst, & Stede (1993) proposed 38 dimensions for differentiating near-synonyms, which were further categorized into semantic and stylistic variations. Stede (1993) focused on the latter and further decomposed them into seven scalable sub-categories. Inkpen & Hirst (2006) organized near-synonym variations into a hierarchical structure, combining stylistic and attitudinal variation into one class in parallel to denotational differences.

Despite the variety of categorization methods, stylistic variation among near-synonyms is an important dimension that has been frequently addressed. In this study, we hypothesize that the stylistic nature of nuances correlates to the degree of difficulty in choosing between near-synonyms. Contrasting some recent studies that focus on contextual preferences of synonyms (e.g., Arppe & Järvikivi 2007), we elect to investigate the internal features of near-synonym nuances. Specifically, we adopt the notion of *core vocabulary* to associate stylistic variation in theory with the difficulty level of near-synonym choice in practice. Core vocabulary consists of "words that suffice to define all of the remaining vocabulary of a language" (Lehmann 1991). It was first related to stylistic variation among near-synonyms by Stede (1993). Carter (1987) listed ten features of core vocabulary, among which, *associationism* is the "bridging" dimension between stylistic variation and core vocabulary. It is characterized by scalable dimensions closely resembling those Stede used for characterizing stylistic variations. Carter claimed that core vocabulary words are relatively neutral on these scales, indicating fewer stylistic variations among them.

Notably, some of Carter's features of CV are readily verifiable using computational linguistic techniques. The collocability of a word, for example, can be approximated by the number of co-occurring word types (normalized by the number of senses to eliminate the confounding factor of polysemy); neutrality in field of discourse can be verified by a word's distribution across different genres in a balanced corpus. Multiple linguistic resources are combined in our study to achieve an empirical characterization of core vocabulary.

To test our hypothesis, a near-synonym lexical choice task (Edmonds 1997) is employed to measure difficulty levels. In this task, lexical gaps are created in sentences from a corpus by removing members of a near-synonym cluster. The sentences are then presented to subjects whose task is to determine from context which member of the cluster is the missing word. Experiments in existing studies have shown great variance in the performance (and hence in level of difficulty) on different near-synonym clusters (Edmonds 1997; Inkpen 2007). Our study shows that such variance is correlated with differing degrees of coreness of the near-synonyms, and in turn, different types of near-synonym variations. Counter to intuition, the seemingly subtle stylistic nuances are usually easier for subjects to distinguish than non-stylistic differences.

Bibliography

Arppe, Antti & Järvikivi, Juhani. 2007: Every method counts: combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory*, 3(2), 131–159.

Carter, Ronald. 1987: *Vocabulary: Applied Linguistic Perspectives*. Allen & Unwin.

Cruse, D. A. 1986: *Lexical Semantics*. Cambridge University Press.

DiMarco, Chrysanne; Hirst, Graeme; & Stede, Manfred. 1993: The semantic and stylistic differentiation of synonyms and near-synonyms. *AAAI Spring Symposium on Building Lexicons for Machine Translation*, 114–121.

Edmonds, Philip. 1997: Choosing the word most typical in context using a lexical co-occurrence network. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 507–509.

Inkpen, Diana. 2007: A statistical model for near-synonym choice. *ACM Transactions on Speech and Language Processing*, 4, 1–17.

Inkpen, Diana & Hirst, Graeme. 2006: Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*. 32, 223–262.

Lehmann, Hubert. 1991: Towards a core vocabulary for a natural language system. *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics,* 303–305.

Stede, Manfred. 1993: Lexical choice criteria in language generation. *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics,* 454–459.