Pragmatic profiling of business corpora: speech act tagging

This presentation describes the first stage of a two-year project which applies corpus analysis and natural language processing techniques to create a comprehensive profile of the pragmatic characteristics of spoken, written, and email Business English. In particular, we discuss the feasibility of extending a speech act tagger developed for workplace emails written by non-native speakers (cf. De Felice and Deane 2009) to corpora of native-speaker Business English, such as the Wolverhampton Business English corpus (10 million words of written text), the Enron email corpus (Berry, Browne, & Signer, 2007), and the Cambridge and Nottingham Spoken Business English Corpus (1 million words).

The speech act tagger has been designed to recognise speech acts typical of email communication such as requests, orders, and commitments. The challenges encountered in applying and adapting the tagger to the spoken and written data highlight how these forms of communication differ from email language, helping us draw a picture of pragmatic variation across the three types, for example in the differing frequencies of particular speech acts, or in the way they are introduced and formulated.

Sentence-level speech act tagging is the first step towards a more detailed analysis of the different types of speech acts, which will consider their lexical and grammatical characteristics, such as which verb forms are most common, or which lexical items feature most often as subjects.

Understanding how speech acts are typically formulated, and how they contribute to the discourse structure of business communication, are key elements for the description of the different forms of Business English. This information is of particular benefit to those unfamiliar with the conventions of this type of language, be they non-native speakers or those just entering the workforce for the first time.

Berry, M., Browne, M., & Signer, B. (2007). *2001 Topic Annotated Enron Email Data Set*. Philadelphia: Linguistic Data Consortium.

Rachele De Felice and Paul Deane (2009). *Identifying speech acts in emails: Business English and non-native speakers.* Corpus Linguistics Conference, Liverpool, UK