Auer, Anita: The Leiden Northern English Letter Corpus: a database to measure the influence of normative grammarians

It is one of the aims of the Leiden-based project *The codifiers and the English language: tracing the norms of Standard English* to measure the influence of normative eighteenth-century grammarians on actual language usage, in particular the language of private letters. To this purpose, *The Leiden Northern English Letter Corpus* is currently being compiled, which covers the period from 1750 to 1900 and is sociolinguistically stratified according to gender, social class, and educational background. This corpus, based on manuscript letters, contains a substantial amount of correspondence of individual writers (their in and out-letters), which allows the study of their language (micro level) as well as the study of the correspondence on a larger scale (macro level).

The purpose of this paper is two-fold: first, I will describe the development and current status of *The Leiden Northern English Letter Corpus*. Second, I will investigate the corpus on a micro level with respect to the development of the inflectional subjunctive (in competition with the indicative form and modal auxiliaries) in adverbial clauses. The results will be compared to the outcome of a multi-genre study (Auer 2006), based on ARCHER (Biber & Finegan 1990-1993/2002/2007), which shows that the trajectory of decline of the inflectional subjunctive was interrupted in the second half of the eighteenth century by an increase of the inflectional form. *The Leiden Northern English Letter Corpus* will enable me to study in sociolinguistic detail the suggestion made by Auer (2006) that this development reflects the short-term influence of normative grammarians.

References

Auer, Anita (2006). "Precept and Practice: The Influence of Prescriptivism on the English Subjunctive". In Christiane Dalton-Puffer & Dieter Kastovsky & Nikolaus Ritt & Herbert Schendl (eds.) *Syntax, Style and Grammatical Norms: English from 1500-2000*. Linguistic Insights. Frankfurt; Bern, etc.: Peter Lang, pp. 33-53.

Biber, Douglas & Edward Finegan, comp. 1990-1993/2002/2007. *A Representative Corpus of Historical English Registers* (ARCHER). Northern Arizona University, University of Southern California, University of Freiburg, University of Heidelberg, University of Helsinki, Uppsala University, University of Michigan, and University of Manchester.

Axelsson, Karin: Tag questions in British fiction

The canonical tag question is a typical feature of spoken British English. It has been found in a recent corpus-based study that there are nine times as many tag questions in colloquial British English as in colloquial American English (Tottie & Hoffmann 2006). Tag questions are, however, also used in the written mode, and then of course primarily in fiction. The language of dialogue in fiction is a genre of its own well worth studying (cf. Oostdijk 1990).

The overall aim of my Ph.D. thesis is to investigate whether tag questions are used in the same way in British fiction dialogue as in natural British conversation. Interesting features are structure of the tag, polarity (reversed or constant), sentence type in anchor (declarative, imperative, interrogative or exclamative), position (appended or inserted), ellipsis in the anchor and, of course, also pragmatic functions.

The data are taken from the British National Corpus: for the fiction study, a thinned sub-corpus restricted to the imaginative domain, book as medium, 1985-1993 as publication date and UK and Ireland as domicile of author. The results of this study will then be compared to the results of a similar study of the spoken part of the BNC, where the context-governed part and the demographic part will be studied separately.

I will present some preliminary findings from this fiction study. There will also be a discussion of some methodological problems, i.e. how to define a tag question and how to establish the proportion of direct speech in the fiction part of the BNC (cf. Semino & Short 2004).

References

Oostdijk, Nelleke. 1990. The language of dialogue in fiction. *Literary and Linguistic Computing*, 5:3, 235–241.

Semino, Elena & Mick Short. 2004. *Corpus stylistics: speech, writing and thought presentation in a corpus of English writing.* London & New York: Routledge.

Tottie, Gunnel & Sebastian Hoffmann. 2006. Tag questions in British and American English. *Journal of English Linguistics*, 34:4, 283–311.

Chao Castro, Milagros: On the origin and use of dual-form adverbs: A corpus-based approach

A dual-form adverb can be defined as an item which derives from an elementary adjective (Ungerer 1988) and which presents two variants, a suffixless and a suffixed adverbial form, e.g. *great*/ *greatly* (Nevalainen 1994a: 248-249). Although adverbs have been studied extensively both for early stages of the language and Present-day English by authors such as Dixon (1982), Donner (1991), Nevalainen (1994a, 1994b, 1997) or Swan (1994), among others, the analysis of these items in the Late Modern English has remained largely overlooked. Therefore, this paper tries to fill part of this gap by investigating the behaviour of dual-form adverbs in the 18th century.

In order to understand the use of these adverbial variants, the word-formation processes involved in their development have been analyzed. Therefore, conversion is used to explain the origin of the suffixless adverb, while derivation by means of the suffix *–ly* justifies, in principle, the appearance of the suffixed form. However, it must be noticed that the analysis of these word-formation processes has revealed the existence of homomorphic adjectives in *–ly* which are often the origin of their adverbial counterparts.

The *Century of Prose Corpus* (COPC), a data base which covers the time span 1680-1780, has been used as a source of data for the analysis. In those cases in which the number of examples found is too low, the *OED* has served as a source of additional evidence. A graphic representation of the evolution and behaviour of some of the dual-form adverbs found in this data base is also offered.

References:

COPC = *The Century of Prose Corpus* (1995). L.T. Milic (compiler). Cleveland: Department of English, Cleveland State University.

Dixon, R. (1982). *Where have all the adjectives gone?* Berlin, New York & Amsterdam: Mouton Publishers.

Donner, M. (1991). "Adverb form in Middle English." *English Studies* 72/ 1: 1-11.

Nevalainen, T. (1994a). "Aspects of adverbial change in Early Modern English." In Kastovsky, D. (ed.). *Studies in Early Modern English.* Berlin: Mouton de Gruyter, 243-259.

OED = The Oxford English Dictionary on CD-ROM (1989). [2nd ed.]. Ed. by John A. Simpson & Edmund S.C. Weiner. Oxford: O.U.P.

Swan, T. (1994). "A note on Old English and Old Norse initial adverbials and word-order with special reference to sentence adverbials." In Swan, T., E. Worck, and O. J. Westwik (eds.). *Language Change and Language Structure. Older Germanic Languages in a Comparative Perspective.* Berlin: Mouton de Gruyter, 233-270.

Ungerer, F. (1988). *Syntax der englischen Adverbialen* (Linguistische Arbeiten 215). Tübingen: Niemeyer.

Chateau, Carmela: Drift and shift: how "continental" and "continents" move in geological English

Kuhn's theory of scientific revolutions suggest that at moments of paradigm shift, scientists can no longer communicate adequately as meaning is in a state of flux. *WebsTerre*, a diachronic corpus of geological English (approximately 2 million words) has thus been designed to investigate language change at a key moment of paradigm shift in the field of Earth science, the shift from a fixed earth paradigm, through continental drift, to the current theory of plate tectonics. The earliest text in the corpus dates from 1830, and the most recent from 1990, but the main corpus is composed of a small central core of key texts and a large collection of articles from the same period (1960-1970).

This report investigates changes in collocation and meaning for "continents" and "continental" in the *WebsTerre* corpus. As a counterpoint reference, the Brown corpus (1 million words) is examined for patterns in general American English from 1961 (the entry stage for the main corpus). The BNC Baby (4 million words) is studied for patterns in usage in British English up to 1991 (the final stage of the main corpus).

This analysis is also set against a text analysis of the same terms in the collection of 17 "eyewitness" reports compiled by Naomi Oreskes in 2001.

References

Dury, Pascaline. 2004. "Building a bilingual diachronic corpus of ecology: The long road to completion". *Icame Journal* 28.5–16.

Goodney, David E. & Carol S. Long. 2003. "The collective classic: A case for the reading of science". *Science & Education* 12.167–184

Jacques, Marie-Paule. 2005. "Pourquoi une linguistique de corpus?". *La Linguistique de Corpus* (*Actes des 2èmes Journées Linguistique de Corpus à Lorient*) ed. by Geoffrey Williams, 21–30. Rennes: Presses Universitaires de Rennes.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Nitecki, Matthew H., J. L. Lemke, Howard W. Pullman & Markes E. Johnson. 1978. "Acceptance of plate tectonic theory by geologists". *Geology* 6.661–664.

Oreskes, Naomi, 2003 *Plate Tectonics: an insider's history of the Modern Theory of the Earth.* Boulder: Westview Press.

Damascelli, Adriana Teresa: Distinctive features in argumentative essays written by Italian and English university students. Comparing ICLE-IT and LOCNESS

The International Corpus of Learner English (ICLE), a project started at Université Catholique Louvain-La-Neuve, Belgium, is a corpus which includes argumentative essays written by university students of English with different language backgrounds, e.g. German, French, Finnish and Italian. The project also provides a comparable corpus of native English essays (LOCNESS) for contrastive analysis. Both corpora are in raw format, i.e. plain text, but can be automatically POS annotated through TOSCA, a tagger which includes a tagset of 270 items.

Although grammatically annotated corpora are common, most grammatical studies have been carried out by focussing on lexical items, for example modals and complementisers, while little attention has been paid to the identification of syntactic patterning.

According to Kennedy (1996) tags should not be considered as isolated items but as sets which appear in sequence and can indicate syntactic structures and, in particular, syntactic patterning. The quantification of co-occurring tags often reveals distinctive features in texts by particular writers. This kind of analysis, which is commonly carried out in *stylometry* for authorship attribution, can be useful for other purposes, for example to explore learners' writing style.

The present study refers to the analysis carried out by Aarts and Granger (1997) who isolate co-occurring POS tags (e.g. PREP ART N, N PREP N) in order to identify stylistic features in the Finnish, French and Belgian components of ICLE and in the LOCNESS. Following the TOSCA ICE-tagset, the Italian component will be explored in order to find three-tag sequences, or trigrams, to identify stylistic features and compare them to those found in the LOCNESS.

In this study I will aim to find out whether the Italian component is characterised by distinctive features shared with other sub-components or whether there are deviations which suggest a different argumentative stylistic model.

References

Aarts, Jan e Sylviane Granger, (1997), "Tag sequences in learner corpora: a key to interlanguage grammar and discorse", in Granger, Sylviane (ed.), *Learner English on Computer*, London: Longman

Felton, R., (1996), "A new procedure for author attribution", in *ALL-ACH '96. Conference Abstract*, Norwegian Computing Centre for the Humanities, University of Bergen, pp.74-75

Granger, Sylviane (ed.), (1997), *Learner English on Computer*, London: Longman

Granger, Sylviane e Paul Rayson, (1997), "Automatic profiling of learner texts", in Granger, Sylviane (ed.), *Learner English on Computer*, London: Longman

Holmes, D., (1994), "Authorship attribution", in *Computers and the Humanities*, vol. 28, pp.87-106

Hoover, David L., (2002), "Frequent Word Sequences and Statistical Stylistics", in *Literary & Linguistic Computing*, vol. 17, n. 2, pp. 157-180

Hoover, David L., (2001), "Statistical Stylistics and Authorship Attribution: an Empirical Investigation", in *Literary & Linguistic Computing*, vol. 16, n. 4, pp. 421-444

Kennedy, Graeme, (1996), "The corpus as a research domain", in Greenbaum, S. (ed.), *Comparing English Worldwide*, London: Clarendon Press, pp.217-226

O'Reilly and Associates, (1991), *UNIX V*, New York: Addison-Wesley

Scott Mike, (1998), WordSmith Tools, vs. 3.0

TOSCA Research Group, (1997), *TOSCA tagger*, vs. 1.1

Degani, Marta: Re-analysing the semi-modal *ought to*

The present paper relies on the assumption that it is hard to draw clear-cut boundaries between the semantic and pragmatic values conveyed by modal expressions in English, particularly when it comes to the discrimination between objective vs. subjective modality. Indeed, as clearly highlighted by Coates (1983) and more recently by a number of studies on the topic (Westney 1995, Papafragou 2000, Warnsby 2006, among others), modality and its linguistic realizations are to be viewed more and more as a fuzzy system affected by 'indeterminacy' and often resulting in phenomena like 'gradience', 'ambiguity' and 'merger'.

Bearing that in mind, the focus of this paper is on the semi-modal *ought to* and my aim will be to analyze how its co-text and context may affect the overall semantic and pragmatic value of the utterance. Indeed, *ought to* has been traditionally considered as pertaining to deontic (or root) modality, expressing a higher degree of objectivity than that conveyed by other modalized patterns like *should*, for example. The working hypothesis of the paper is that *ought to* has been gradually moving towards the cline of (inter)subjectivity (Traugott et al. 2002) to the point that it now frequently overlaps with the values traditionally conveyed by *should* both in epistemic and in non-epistemic contexts. The data will be retrieved from the so-called 'Brown family' of corpora (*Brown*, *Frown*, *LOB* and *FLOB*), so as to identify any possible short-term diachronic changes within the British and American varieties under scrutiny.

References

Athanasiadou, A. & C. Canakis, B. Cornillie (eds.), (2006), *Subjectification: Various Paths to Subjectivity*, Berlin – New York, Mouton de Gruyter.

Coates, J. (1983), *The Semantics of the Modal Auxiliaries*, London, Croom Helm.

Langacker, R.W. (2003), "Extreme Subjectivisation. English Tense and Modals". In H. Cuyckens et al. (eds.) *Motivation in Language*. Amsterdam/Philadelphia: John Benjamins

Mackenzie J. Lachlan & M. de los Angeles Gómez Gonzáles (eds.), (2004), *A New Architecture for Functional Grammar*, Berlin-New York, Moutone de Gruyter.

Myhill, J. (1997), "*Should* and *ought*: the rise of individually oriented modality in American English". *English Language and Linguistics* 1(1): 3-23.

Nordlinger, R. & E. Closs Traugott (1997), "Scope and the development of epistemic modality: evidence from *ought to*". *English Language and Linguistics* 1(2): 295-317.

Palmer, Frank R. (2001), *Mood and Modality*, 2nd ed. Cambridge, Cambridge Textbooks in Linguistics (1st ed. 1986).

Papafragou, A. (2000), *Modality: Issues in the Semantics and Pragmatics Interface*. Amsterdam: Elsevier.

Stein, D. & S. Wright (eds.), (1995), *Subjectivity and Subjectivisation: Linguistics Perspectives*, Cambridge, CUP.

Traugott, E.C. & R. Dasher (2002), *Regularity in Semantic Change*, Cambridge, CUP.

Warnsby, A. (2006), *(De)coding Modality*: *The Case of* Must, May, Maste, *and* Kan. Lund Studies in English 113. Lund: Lund University.

Westney, P. (1995), *Modals and Periphrastics in English*. Tubingen: Max Niemeyer.

Estling Vannestål, Maria: *Only time will tell*: Phraseological patterns in native speakers' and learners' use of expressions with the noun *time*

As observed by Pawley & Syder (1983) more than two decades ago, even advanced language learners tend to sound unidiomatic in their target language because they do not fully master its phraseology. It has been advocated that teachers pay more attention to the teaching of collocation (e.g. Lewis 2000; Nation 2001), but Granger points to the need for empirical studies to know what and how to teach (1998). For instance, investigations of learners with a particular language background can help teachers of English decide what particular problems their students may have because of different phraseological patterns in L1 and L2.

This presentation outlines a pilot study of phraseological patterns including *time*, which is the most frequent noun in the British National Corpus (BNC). Stubbs (2004) suggests that many words are frequent because they occur in frequently used phrases. Mahlberg (2005) further observes that "time nouns occur in a number of patterns that illustrate different facets of time meanings'. The aim of this particular study is to compare native speakers' and Swedish learners' use of such patterns. Fletcher's PIE program (from 2003/4) was used to extract so-called n-grams including *time* from the BNC. N-grams that were considered to form a phraseological structure were then compared to phrases including *time* in the USE (Uppsala Student English) corpus.

The first part of the study showed (not surprisingly) that many of the phrases found in the BNC were entirely absent from the learner corpus. Another observation is that many phrases that occurred in both corpora were much more frequent in the learner corpus than in the native speaker corpus. The second part of the analysis of the phrases and their co-text further revealed some more subtle phraseological differences between native speakers' and learners' usage.

References

Fletcher, William. 2003/4. PIE: Phrases in English. http://pie.usna.edu.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: collocations and formulae. In *Phraseology. Theory, analysis and applications*. Cowie, A. (ed.). Oxford: Clarendon Press. (145-160)

Lewis, Michael (ed.). 2000. *Teaching collocation. Further developments in the lexical approach*. Boston: Thomson & Heinle.

Mahlberg, Michaela. 2005. *English general nouns: A corpus theoretical approach*. Philadelphia: John Benjamins.

Nation, Paul. 2001. *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Pawley, A. & F. Syder. 1983. Two puzzles for linguistic theory. In J. Richards and R. Schmidt (eds.). *Language and communication.* London: Longman. (191-226)

Stubbs, Michael. 2004. On very frequent phrases in English: distributions, functions and structures. A revised version of a plenary lecture given at *ICAME 25*, in Verona, Italy, 19-23 May 2004 http://www.uni-trier.de/uni/fb2/anglistik/Projekte/stubbs/icame-2004.htm

Fairon, Cedrick: Automatizing the collection of specialized corpora from RSS feeds

This paper presents a new approach and software for collecting specialized corpora on the Web. This approach takes advantage of a very popular XML-based norm used on the Web for sharing content among websites: RSS (Really Simple Syndication). Nowadays many of the press groups around the world offer RSS-based news feeds on their Web sites which allow easy access to the recently published news articles (see for instance http://www.nytimes.com/services/xml/rss/index.html). Blogs and forums may also offer RSS access to their content. RSS sources are very diverse: it is possible to find specific feeds on many different themes, in many languages and for all kinds of text genre.

After a brief introduction to RSS, we explain the interest of this type of data sources in the framework of corpus development and we present Corporator, Open Source software which was designed for collecting corpora from RSS feeds. Efforts have recently been undertaken to make the software easier to use and more robust with improved performance at filtering HTML and retrieving articles split on several Web pages. A short evaluation will be proposed.

References

Fairon, Cédrick. 1999. Parsing a Web site as a corpus. In C. Fairon (ed.). *Analyse lexicale et syntaxique: Le système INTEX*, Lingvisticae Investigationes Tome XXII (Volume spécial). John Benjamins Publishing, Amsterdam/Philadelphia, pp. 327-340.

Fairon, Cédrick. 'Corporator: A tool for creating RSS-based specialized corpora'. In Proceedings of the Workshop Web as corpus. EACL 2006. Trento.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, Vol. 29(3): 333-348.

Renouf, Antoinette. 1993. 'A Word in Time: first findings from the investigation of dynamic text'. In J. Aarts, P. de Haan and N. Oostdijk (eds), *English Language Corpora: Design, Analysis and Exploitation*, Rodopi, Amsterdam, pp. 279-288.

Renouf, Antoinette. 2003. 'WebCorp: providing a renewable energy source for corpus linguistics'. In S. Granger and S. Petch-Tyson (eds), *Extending the scope of corpus-based research: new applications, new challenges*, Rodopi , Amsterdam, pp. 39-58.

Faya, Fátima: On the competition of the courtesy markers *please, pray* and *if you please* in the period 1850-1950: evidence from ARCHER

Across languages requests are typically attenuated by means of so-called 'courtesy markers' (Quirk *et al*. 1985: §8.90) serving a mitigating function, such as present-day English *please* (Blum-Kulka *et al*. 1989: 281-283). Such items tend to be replaced cyclically since they wear down with frequent use and experiment a need to be reinforced. In the last two centuries three courtesy markers have been in competition in English, namely *please*, *pray* and *if you please*. Whereas *please* is nowadays the default marker in requests, the usage of *pray* and *if you please* in present-day English is highly restricted, if not archaic. In fact *pray* is regarded as "especially literary or old use" (*CIDE* s.v. *pray* adv.) or "old use or ironic" (*OALD* s.v. *pray* adv.) and *if you please* "old-fashioned" (*OALD* s.v. *please* v.) or "formal dated" (*CIDE* s.v. *please* v.). A search in the BNC shows the frequencies of *pray* (0.68) and *if you please* (0.89) are extremely low compared to that of *please* (131.15)[1] The aim of this paper is to give an account of the distribution of *please*, *pray* and *if you please* during a time-span of 100 years, taking the middle of the 19th century as a starting point, since the use of *please* was not frequent before this period (Tieken and Faya forth., Faya forth.). I will pay attention to the following issues, among others: (i) frequency of the three items in order to pinpoint the time at which the present-day restrictions in the use of *pray* and *if you please* began to hold; (ii) text-types favouring their selection; (iii) whether or not these markers differ as regards their pragmatic functions. For these purposes I will use data from ARCHER, since it covers the period chosen and enables a multi-genre approach.

[1] Searches for *pray* and *please* are lemma queries as "adverb" and for *if you please* a standard query. Frequencies per million words.

References

Blum-Kulka, Shoshana, Julianne House and Gabrielle Kasper (eds.). 1989. *Cross-Cultural pragmatics: Requests and Apologies*. Norwood: N.J. Ablex.

BNC = The British National Corpus, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. <http://www.natcorp.ox.ac.uk/>

CIDE = Procter, Paul (editor-in-chief). 1995. *Cambridge International Dictionary of English*. Cambridge: Cambridge University Press.

Faya Cerqueiro, Fátima María. Forthcoming. "*Please say what you mean*: Origin and position of the courtesy marker *please* in the nineteenth century." In *Proceedings of the 29th International AEDEAN Conference*.

OALD = Hornby, Albert Sydney. 2000 [1948]. *Oxford Advanced Learner's Dictionary of Current English*. Sally Wehmeier (ed.). Oxford: Oxford University Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Tieken-Boon van Ostade, Ingrid and Fátima María Faya Cerqueiro. Forthcoming. "Saying 'please' in Late Modern English." In *Of Varying Language and Opposing Creed: New Insights into Late Modern English*. Javier Pérez-Guerra *et al*. (eds.). Bern: Peter Lang.

Forchini, Pierfranca: "I mean, what was that about?" Spontaneous and non-spontaneous conversation compared.

Spoken language is a variety characterized by specific linguistic features (e.g. interjections, backchannels, attention signals, repetitions, reformulations, hesitators, discourse markers, vocatives, *inter alia*, cf. Halliday 1985, Biber *et al*. 1999, McCarthy 2003) that clearly distinguish it from written language. The principle aim of the present study is to investigate whether the characteristics of spontaneous conversation are also to be found in movie language, an instance of non-spontaneous, prefabricated speech which is "written to be spoken as if it were not written" (cf. Gregory 1967 and Nencioni 1976) in order to sound authentic; and if so, to what extent.

Within the framework of a wider investigation of spontaneous vs non-spontaneous linguistic devices (such as *discourse markers*, *hedges*, and *interjections*), the present paper focuses in particular on *I mean*, and investigates its frequency of occurrence, semantics and pragmatics in a corpus of transcripts of dialogues from contemporary American movies directed from 2000 on, as compared to the US spoken subcorpus of the *Bank of English*.
A further aim is to analyze the translations of *I mean* in the Italian dubbed versions of the movies examined in order to highlight the functional, rather than lexical, nature of discourse markers (cf. Bazzanella and Morra 2000) and to verify whether they provide evidence of universal features of translation behaviour through multimedia translation studies (cf. Baker 1998, Ulrych 1998) or whether the linguistic choices are influenced by the languages involved.

The analyses are corpus-driven (cf. Francis 1993, Tognini-Bonelli 2001) in that the theory is built up in the presence of the evidence found in the US spoken subcorpus of the *Bank of English* and in the corpus of transcripts of American movies and their dubbed Italian versions.

References

Baker, Mona (ed.). 1998. *Routledge encyclopaedia of translation studies*. London: Routledge.

Baker, Mona, Francis, Gill, Tognini-Bonelli, Elena. 1993. *Text and Technology: Essays in Honour of John Sinclair*. Amsterdam: John Benjamins.

Bazzanella Carla, Lucia Morra. 2000. "Discourse markers and the indeterminacy of translation", in Iørn Korzen, Carla Marello (eds.) *Argomenti per una linguistica della traduzione, On linguistic aspects of translation, Notes pour une linguistique de la traduction*. Alessandria: Edizioni dell'Orso.

Biber, Douglas, *et al*. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Francis, Gill. 1993. 'A corpus-driven approach to grammar: principles, methods and examples' in Baker *et al*. (eds.) 137-156.

Gregory, Michael. 1967. "Aspects of Varieties Differentiation", in *Journal of Linguistics*, 3, pp. 177-98.

McCarthy, Michael. 2003. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

Nencioni, Giovanni. 1976. *Di scritto e di parlato. Discorsi linguistici*. Bologna, Zanichelli.

Pavesi, Maria. 2005. *La traduzione filmica. aspetti del parlato doppiato dall'inglese all'italiano*. Roma: Carocci.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam, Philadelphia: John Benjamins.

Halliday, Michael. 1985. *Spoken and Written Language*. Victoria: Deakin University.

Ulrych, Margherita. 1998. "Locating universal features of translation behaviour through multimedia translation studies", in Bollettieri Bosinelli, R. Maria (eds.). *La traduzione multimediale: quale traduzione per quale testo? Atti del convegno internazionale: La traduzione multimediale*: Bologna: CLUEB.

Gesuato, Sara: 'Coming to know' and 'coming to do': semantic and syntactic restrictions on the aspectual usage of a motion verb

The verb *COME* followed by an infinitive signals goal-oriented motion (with *to* paraphrasable as 'so as to'; e.g. *I came to ask you what you wanted me to do*; *the people who come to be healed*) or gradual completion of a process (with *come* paraphrasable as 'end up (V-ing)'; e.g. *he came to believe something strange had happened; It came to be known as the finger issue*). In the former case, *COME* is used literally, the infinitive encoding a clause of purpose whose verb denotes deliberate actions (e.g. *dance*, *lecture*). In the latter, *COME* is used aspectually, the infinitive encoding a catenative complement, whose verb denotes involuntary experiences (e.g. *dread, be built*).

Examples from the Bank of English (about 4,300; 79% from written sources) show that the aspectual sense of *COME* ('end up': 64%) is more common than the literal one ('get closer': 28%), but that ambiguity arises (8%) when *COME* expresses the instantaneous result of a process, which makes it compatible with an agentive ('decide') or (externally) causal ('happen') interpretation (e.g. *how did you come to be setting up a business like this*). The data also shows that the construction combines with auxiliaries and morphological markers to encode temporal-aspectual distinctions (past (31%), perfect (20%), simple present (32%), future (3%), progressive (7%)) and modality (7%), but that restrictions apply (e.g. the progressive may occur only once in the construction: *Femininity itself was coming to be perceived rather differently*; *that's why I come to be paying that amount*; or an auxiliary occurs in the progressive matrix clause only if this encodes 'motion': *his neighbour will be coming to pick him up*).

The data suggests that the resultative notion conveyed by the *COME* + infinitive construction ('come about') is typically associated with the presentation of whole, complete events seen as 'unintentional/unwanted consequences'.

Kaislaniemi, Samuli: Transcribing historical correspondence: Methodological considerations for corpus compilers

Compilers of historical corpora have started to stress the importance of the true 'witness' or 'informant' that is the original text (Lass 2004; Dury 2006). In order to retain textual accuracy, the "main desiderata" for historical corpora are, as Lass puts it, "maximal retention of historical information and minimal addition, with the additions all removable if necessary" (2004: 45–6). That is, historical corpora should be compiled from original sources or impeccable diplomatic transcripts in order to ensure that studies are conducted on 'uncontaminated' texts. Furthermore, any tagging of the texts should be minimally interpreting, and diplomatic transcriptions of the original texts should be viewable by users of the corpus. Just which features of the texts — grammatical, spatial, and semiotic (e.g. plurals, spaces and fonts/hands) — should be tagged into the corpus, and how they should be encoded, remain topics for discussion and development (cf. Lass 2004: 42–5; Dury 2006).

My PhD project is an electronic edition of the early correspondence of Richard Cocks, English merchant (1600–1610). The project is an indirect offshoot of the *Corpus of Early English Correspondence*, an experiment in historical sociolinguistics applied to corpus compilation (see, for instance, Nevalainen & Raumolin-Brunberg 2003). My thesis attempts to combine the requirements of electronic editions and linguistic corpora, of historians and linguists. It is also a part of the continuing search for standards in the digitization of cultural material — a broad field which includes linguistic corpora (cf. the Text Encoding Initiative guidelines at http://etext.lib.virginia.edu/standards/tei/teip4/)

This paper reports on some methodological and practical solutions in creating the edition — responses to the challenge of retaining both readability and the possibility to conduct linguistic searches while encoding extralinguistic features.

References

Dury, Richard. 2006. "A corpus of nineteenth-century business correspondence: Methodology of Transcription". *Business and Official Correspondence: Historical Investigations*, ed. by Marina Dossena & Susan M. Fitzmaurice. Bern: Peter Lang, 193–205.

Lass, Roger. 2004. " 'Ut custodiant litteras': Editions, corpora and witnesshood". *Methods and Data in English Historical Linguistics*, ed. by Marina Dossena & Roger Lass. Bern: Peter Lang, 21–50.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Pearson Education. (Longman Linguistics Library).

Kanoksilapatham, Budsaba: Subject-verb agreement produced by Thai University Students

Despite great efforts in developing English education in Thailand, the nationwide study conducted by the Office of the National Education Commission (2000) showed that the achievement of Thai learners was unsatisfactory. In response to the impact of globalization, teaching strategies and learning potential in English classes need to be improved to satisfy the demands of the international community. It is inevitable that learners of any language make errors in the process of language learning. Therefore, error analysis is deemed significant theoretically and practically, revealing how a language is learned and how errors are corrected. Among all types of errors, subject-verb agreement was the most frequent one in Thai university students' writing (Pongsiriwet, 2001). Given the fact that this type of errors is so pervasive, this study focused on the errors of it in 53 argumentative essays that were at least 500 words and were composed by Thai university students. The identified errors were further classified into six categories. Interestingly, most of the errors found in the corpus were those that occurred in standard sentence structure, and most of the subjects in these standard sentences were nouns. The findings elucidate how Thais learn English and what factors possibly contribute to these errors. The implications of the findings are useful in providing a basis for improving instruction concerning the subject-verb agreement and for designing the English curriculum, allowing Thais to succeed in learning English.

Kübler, Natalie: The Learner-Translator Corpus

This paper deals with the issue of creating and exploiting a Learner Translator Corpus (LTC) for applications such as translation training and translation aids. The LTC is part of a European project, Multilingual e-Learning in LANGuage Engineering (MeLLANGE).

The MeLLANGE consortium collected translations made by students from and into English and several other European languages (Fr, CA, DE, ES, IT); an error typology was developed in parallel, in order to allow the project's members to annotate the students' translations. This presentation will show according to which conventions the corpus was annotated. The issue of analysing and exploiting the corpus will then be tackled, and some results and example of use will be demonstrated.

Website: http://mellange.eila.jussieu.fr/

Ljung, Magnus: Pragmatics in British swearing

The last decade has seen a growing number of publications on swearing, in particular swearing in English, for example Hughes 2006, 1998, McEnery 2005, McEnery and Xiao 2003, 2004, and – from a different perspective - van Lancker and Cummings (1999).

Most of these studies describe swearing in terms of functional categories like expletives, oaths, emphatic adverbs etc., categories that operate in terms of a limited number of "swearwords", a subset of "bad language" containing vernacular terms for activities or things once considered "taboo" like the sexual act, the sex organs, excrement.

An aspect of swearing that has received less attention is how close interjections like *Shit*!, *Fuck*! etc. are to the category of pragmatic markers. Like these they may be used to express speaker attitudes, to signal the organization of text and to deliver interactional signals of various kinds. In fact, these interjections meet most of the widely accepted criteria for pragmatic markers proposed in Brinton (1996:33 ff.) and may be regarded as the output of the same processes of grammaticalization commonly supposed to underlie all pragmatic markers (cf. e.g. Hopper and Traugott 1993).

In the completed study that will eventually result from my investigation I will argue that there are good reasons to include interjections involving swearing among the pragmatic markers. In my ICAME talk I will focus on the less time-consuming task of attempting to analyze the pragmatic functions of a small set of interjections found in a one-million-word corpus containing texts from the spoken component of the BNC, and pointing to the difficulties involved in such an analysis (for which see e.g. Andersen 1999:15 ff. and Stenström 2006).

References

Gisle Andersen 1999 *Pragmatic markers and sociolinguistic variation*: *a corpus-based study*. Bergen: University of Bergen

Laurel J. Brinton 1996. *Pragmatic markers in English*: *grammaticalization and discourse Functions*. Berlin: Mouton de Gruyter.

Geoffrey Hughes 2006 *An Encyclopedia of swearing*. Armonk N.Y.:Sharpe.

Geoffrey Hughes 1998 *Swearing*: *a social history of foul languages, oaths and profanity in English.* Oxford:Blackwell.

Magnus Ljung *Svordomsboken*. Stockholm: Norstedts Akademiska Förlag.

Tony. McEnery 2005 *Swearing in English*.Abingdon:Routledge.

Tony McEnery and R.Z. Xiao 2004 "Swearing in modern British English: the case of FUCK in the BNC" *Language and Literature* 2004 13:235-268

Tony McEnery and R.Z.Xiao "Fuck revisited" *Corpus Linguistics* 2003 28.31.

Anna-Brita Stenström 1994. *An introduction to spoken interaction*. London:Longman

Anna-Brita Stenström. 2006. "Taboo words in teenage talk. London and Madrid girls' conversations compared". *Spanish in Context* 3 2006 (1) 115-138.

D.van Lancker and J.L.Cummings 1999 "Expletives: neurolinguistic and neurobehavioural perspectives on swearing. *Brain Research Reviews* 31:83-104.

Mäkinen, Martti & Stenroos, Merja: The Middle English Grammar project – working towards a corpus of Middle English localisable texts

This paper will report work in progress. The Middle English Grammar Project, a joint project ongoing at the Universities of Glasgow and Stavanger, is aiming at publishing a grammar of Middle English which will eventually replace Jordan's *Handbuch der mittelenglischen Grammatik: Lautlehre* (Jordan 1925). The new grammar will cover orthography, phonology and morphology.

The grammar will be based on analyses of Middle English texts in extracts. For this task we have compiled A Corpus of Middle English Localizable Texts. The Corpus will contain about 1,000 texts from the late Middle English period in 3,000-word tranches. All the texts have been localised in *A Linguistic Atlas of Late Mediaeval English* (McIntosh *et al*. 1986). The texts in the Corpus are transcriptions from either the original manuscripts or good-quality microfilms. We have used the same conventions for transcription as have been adopted for *A Linguistic Atlas of Early Middle English* (Laing and Lass under preparation) as we have designed the corpus to be compatible with that resource.

We have scheduled the Corpus to be published from August 2007. The first phase of the publication will provide the research community with an open access Internet site where the transcribed texts that are not constrained by copyright issues can be browsed and searched. In a later phase, a restricted version of the Corpus containing database functionality and/or XML coded texts and an appropriate software will accompany the published texts.

References

http://www.arts.gla.ac.uk/sesll/englang/ihsl/projects/MEG/MEG.htm

Black, Merja, Simon Horobin and Jeremy Smith 2002. 'Towards a new history of Middle English spelling.' In P. J. Lucas and A.M. Lucas (eds), *Middle English from Tongue to Text.* Frankfurt am Main: Peter Lang, 9-20.

Horobin, Simon and Jeremy Smith 1999. 'A Database of Middle English Spelling.' *Literary and Linguistic Computing* 14: 359-73.

Jordan, Richard 1925. *Handbuch der mittelenglischen Grammatik*. Heidelberg: Winter's Universitätsbuchhandlung.

McIntosh, Angus, M.L. Samuels and Michael Benskin 1986. *A Linguistic Atlas of Late Mediaeval English.* Aberdeen: University Press.

Stenroos, Merja 2004. 'Regional dialects and spelling conventions in Late Middle English: searches for (th) in the LALME data.' In M. Dossena and R. Lass (eds), *Methods and data in English historical dialectology*. Frankfurt am Main: Peter Lang: 257-85.

Laing, Margaret, and Lass, Roger (under preparation). *A Linguistic Atlas of Early Middle English.*

Minugh, David: The Nixon Idiom

This paper will explore the use of idioms and idiomatic phrases in the Watergate tapes. Recorded in the period during and subsequent to the government-sponsored break-in into the Democratic Party headquarters in 1972, these tapes were subpoenaed by the Watergate Special Prosecution force and subsequently deposited in the US National Archives. They detail the private discussions of President Nixon and his immediate counsellors and comprise some 60 hours of taped conversations. The primary research question is how the social bonding of this small group of men is reflected in greater conformity in above all the informal idiomatic language they used in their closed meetings in the Oval Office. Who first contributes the idiom? Is it taken up by the others in the same or subsequent meetings?

Reference

Cutting, Joan. 2002. "The In-Group Code Lexis", in *Hermes, Journal of Linguistics* (28): 59-80.

Nevalainen, Terttu; Jukka Tyrkkö & Matti Rissanen: A Corpus Resource Database (CoRD)

Despite the number of websites put up by individual corpus linguistics projects as well as those listing corpora and corpus tools, there is to date no focussed online resource through which corpus compilers could publicize information about their corpora. In particular, there is no resource which would allow a researcher or a student not only to find whether a corpus relevant to a given research question is available, but also to learn what the background and composition guidelines of a corpus are, what information is coded in it and what research has been conducted using it. To answer these questions, the VARIENG Research Unit is initiating an online corpus resource, tentatively entitled *A Corpus Resource Database* (CoRD).

The database will not distribute corpora as such, but can serve as a point of contact between users of the service and corpus developers by providing useful information about corpora, including their background, reference details and availability. The CoRD will be an open-access online resource on which academic corpus compilers can make available this basic information about their corpora. The general structure of the database will consist of an open-ended number of modules, each describing a corpus following a set format. These corpus modules will be made available using an online content management tool developed at the University of Helsinki, which supports both text and multimedia content. We will illustrate the *Corpus Resource Database* by discussing the information compiled for it on the Helsinki Corpus of English Texts.

Renouf, Antoinette & Jay Banerjee: Repulsion: How far from Word A to Word B?

In our current research, we propose that there is an unexplored 'force' in text that we call 'repulsion' (Renouf & Banerjee, 2007), which operates in an opposing way to that of lexical collocation (Firth, 1957). By 'repulsion', we mean the intuitively-observed tendency in conventional language use for certain pairs of words **not** to occur together. The goal of our large-scale study is to establish how repulsion operates in text and whether it has the status of an objective and measurable 'force'.

In this presentation we report on the latest progress made in the project, building on our previous reports of the repulsion found specifically between sense-related word pairs (Renouf & Banerjee, forthcoming 2007). Here, we focus on methodologies developed to measure the 'active distancing between words' that differentiate 'repulsion' from routine 'indifference' (non co-occurrence of two un-associated words), and discuss the results obtained. We consider the combined collocate space of a pair of sense-related words, and measure the distances (Davies, unpublished) between each word and all the collocates within the collocate space; and then rank the distances in preference to word A or word B of the pair. We search for patterns/groups that each word in a sense-related pair repels, and attempt to understand the semantic and other factors hidden within the complex phenomenon.

References

Davies, Paul (unpublished). 'Methods of measuring 'distance' between two sense-related words'.

Firth, J. R. (1957). 'Modes of meaning'. In: *J. R. Firth: Papers in Linguistics 1934-1951*. (pp. 190-215). London: Oxford University Press.

Renouf, A.J. and J. Banerjee (2007). 'The Search for Repulsion: a new corpus analytical approach' in eVARIENG: *Methodological Interfaces* as online *Proceedings of 27th International ICAME Conference*, May 2006, Hanasaari, Finland.

Renouf, A.J. and J. Banerjee (forthcoming, 2007). 'Lexical repulsion between sense-related pairs' in M.Mahlberg (ed.) *International Journal of Corpus Linguistics* 12.3. Amsterdam: John Benjamins Publishing Company.

Renouf, A.J. and J. Banerjee (forthcoming). 'The Phenomenon of Repulsion in text' in Leclère, C. et al (eds.) *Special Edition of Proceedings of 25th International Conference on Lexis and Grammar*, Palermo, Sicily, Sept. 6-10, 2006, Lingvisticae Investigationes, Amsterdam: John Benjamins Publishing Company.

Yamazaki, Shunji: Comparative analysis of comparison of adjectives in Modern English

Past studies of the comparison of adjectives in late middle and early modern English (Kytö 1996, Kytö and Romaine 2000, Suematsu 2004) and in modern English (Quirk *et al*. 1985, Leech and Culpeper 1997, Lindquist 2000) have demonstrated the robust use of inflectional comparison with monosyllabic adjectives, whereas trisyllabic or longer adjectives tend to take periphrastic comparison, and disyllabic adjectives exhibit variation between the two forms. Those earlier studies commonly stress several factors that are held to decide the preference between inflectional or periphrastic comparison, such as 1) the number of syllables in the adjective, 2) whether the final syllable is stressed, and 3) what sort of suffix the adjective might have. The present research investigates the effects of adjective length and suffixes on the preference for one form rather than another in different text categories in four corpora of modern English (the Brown, LOB, Frown, and FLOB Corpora) which allow a comparison between British and American English, and an indication of change between 1961 and 1991.

References

Kytö, M. 1996. " 'The Best and Most Excellent Way': The Rivalling Forms of Adjective Comparison in Late Middle and Early Modern English", in J. Svartvik (ed.), *Words: Proceedings of an International Symposium: Lund, 25-26 August 1995*. Stockholm: Almqvist and Wiksell. 123-41.

Kytö, M. and S. Romaine. 2000. "Adjective Comparison and Standardisation Processes in American and British English from 1620 to the Present", in *The Development of Standard English 1300-1800*. Cambridge. 171-94.

Leech, G. and J. Culpeper. 1997. "The Comparison of Adjectives in Recent British English", in T. Nevalainen and L. Kahlas-Tarkka (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique. 353-73.

Lindquist, H. 2000. "*Livelier or more lively?* Syntactic and Contextual Factors Influencing the Comparison of Disyllabic Adjectives", in J. M. Kirk (ed.), *Corpora Galore: Analyses and Techniques in Describing English*. Amsterdam: Rodopi. 125-32.

Suematsu, N. 2004. "The Comparison of Adjectives in 18th-Century English", in *NOWELE: North-Western European Language Evolution*. Denmark: Odense.

Zipp, Lena: Intra- and inter-varietal testing for differences in verb complementation patterns – First explorations of ICE Fiji

In order to build yet another case for the significance of lexico-grammatical features in the description of national varieties of English, this paper investigates verb complementation patterns in written Indo-Fijian English. It introduces my ongoing PhD research on exo- and endonormative models in Fiji, reporting briefly on the current state of the art concerning the compilation of the Fiji component of the *International Corpus of English* (ICE) along with some methodological problems encountered (e.g. author eligibility), and presents preliminary results of intra- and intervarietal testing using the written samples of ICE Great Britain, ICE New Zealand and ICE India as comparable databases.

The question of norm development is crucial for English as a second language (ESL) varieties. According to Schneider's (2003) dynamic model, it is assumed that New Englishes in general follow a cycle of evolution. This leads on the one hand to younger codified first language varieties being accorded norm-providing status (resulting in a pluricentric model of English) and on the other hand to ESL varieties progressing towards endonormative stabilization through a "nativization" phase, for which lexico-grammatical features have been noted to be some of the most productive indicators.

With ICE Fiji subdivided into two parts mirroring the two major ethnic groups of the population (and their respective first languages), the study proposes to examine whether Indo-Fijian English displays differences in individual verb complementation patterns (e.g. *provide for/to/with/ø*) when compared to its ethnic counterpart, Fijian English, or whether it is justified to speak of common structural nativization processes of a single national variety of English in Fiji. An extension of the analysis will position Fiji English in relation to its possible exonormative prestige models: the varieties of three countries with political, economic or cultural influence on Fiji – Great Britain, New Zealand and India.

References

Mukherjee, Joybrato and Sebastian Hoffmann. 2006. "Describing verb-complementational profiles of New Englishes. A pilot study of Indian English". *English World-Wide* 27:2. 147-173

Olavarría de Ersson, Eugenia and Philip Shaw. 2003. "Verb complementation patterns in Indian Standard English". *English World-Wide* 24:2. 137-161

Schneider, Edgar W. 2003. "The dynamics of New Englishes: from identity construction to dialect birth". *Language* 79:2. 233-281

Schneider, Edgar W. 2004. "How to trace structural nativization: particle verbs in world Englishes". *World Englishes* 23:2. 227-249

Tent, Jan and France Mugler. 1996. "Why a Fiji Corpus?" in: *Comparing English Worldwide. The International Corpus of English.* Sidney Greenbaum (ed.). Oxford: Clarendon. 249-261

Zumstein, Franck: The evolution of word-stress variation: from Jones's English Pronouncing Dictionary (1963, 12th ed.) to Wells's Longman Pronunciation Dictionary (1991, 1st ed.)

In the 1960s, Lionel Guierre, who was at the head of a research team in Paris, digitized the twelfth edition of Daniel Jones's *English Pronouncing Dictionary* (hence EPD12) on which he based his study of word-stress. He enriched the corpus with various stress and phonemic variation codes which he described in a paper written in 1966 with the aim of undertaking a comprehensive study of such variations. Yet, he never had the time to achieve this goal. In 1991, Guierre managed to get the original computerized file used to print the first edition of John Wells's *Longman Pronunciation Dictionary* (hence LPD1). He also added several types of lexical information at each word entry such as grammatical category, syllable count, main stress pattern and reverse spelling. Unfortunately, he did not include the variation codes found in the electronic file of EPD12.

Both corpora exist as text-only files from which data-retrieval programs can be launched with the help of software called Macintosh Programmer's Workshop (hence MPW). It has thus been possible to sort out exhaustive lists of words with stress variation thanks to Guierre's codes in EPD12, Guierre's added information in LPD1, and the possibility of setting up many variables in MPW's search commands.

This presentation aims first at giving an account of the work that has been carried out with the above-mentioned data-retrieval processes. Then it proceeds to analyse the evolution of some word-stress variations in the lapse of the 30 years that separate the two dictionaries, such variations being the result of conflicting word-stress rules.

References

Duchet, Jean-Louis 1994. *Code de l'anglais oral*. Paris: Ophrys.

Gimson, Alfred Charles 1975. *An Introduction to the Pronunciation of English*. London: Arnold.

Fudge, Eric 1984. *English Word Stress*. London: Allen & Unwin.

Guierre, Lionel 1979. *Essai sur l'accentuation en anglais contemporain*. PhD dissertation. Paris: Université de Paris VII.

Guierre, Lionel 1984. *Drills in English Stress Patterns*. Paris: Armand Colin-Longman.

Jones, Daniel 1957, *An Outline of English Phonetics*. Cambridge: Heffer.

Jones, Daniel 1963. *English Pronouncing Dictionary*. 12th edition. London: J.M. Dent.

Walker, John 1797, *A Critical Pronouncing Dictionary and Expositor of the English Language*. London: G. G. & J. Robinson, Paternoster-Row; and T. Cabell, Junior, and W. Davies in the Strand.

Fowler, Henry Watson 1996. *The New Fowler's Modern English Usage*. Burchfield R. W. (ed.). Oxford: O.U.P.

Wells, John 1982. *Accents of English*. 3 vols. Cambridge: C.U.P.

Wells, John 1990. *Longman Pronunciation Dictionary*. 1st edition. Harlow: Longman.