Nouvelles approches du corpus en linguistique anglaise / New approaches to corpus in English linguistics

9-10 juin 2016
Avignon
France

Table des matières

ABSTRACT

Widdowson (2004) states that Critical Discourse Analysis (CDA) is biased in as much as it 'cherry picks' the linguistic features it seeks to examine within a dataset. With Fowler (1996), he argues that data selection is fragmentary in CDA because meanings are put on texts rather than taken from texts. However, with the aid of Corpus Linguistic (CL) tools, it becomes possible to interrogate large corpora, extracting hundreds or thousands of text samples in their micro- and macro-contexts(Van Dijk, 2004) for analysis. Thus, CL tools can yield greater data representativenessthanCDA and be used alongside CDA to generate more robust analysis (Stubbs, 1997; Taylor and Marchi, 2009).

This presentation reports part of my work on the discourse representation of Poverty and Social Exclusion(PSE) in British Annual Conferences of the Conservative and Labour Parties, across the 20th and 21st centuries. It focuses on its empirical design, specifically on the steps needed to select a relevant and manageable dataset, from a large corpus of 1 million words, for CDAexamination. I draw upon CL tools (Anthony, 2014; Baker et al., 2013; Kilgarriff, et al., 2004, Baker, 2004) to develop a three-stage analytic procedureto extract and downsize representative data forCDAanalysis.These stages are: (1) examiningSeed Words (SWs) (search words) that refer to PSE in context, extracting their concordance lines, (2) analyzing the word sketches of the SWs in their extracted concordances, grouping their collocates into semantic categories, (3) using (1) and (2) to determine common discourse types of PSE in the corpus.

The application of these stages led to downsizing PSE concordances from an initial list of 6,130 concordances to a final one of 749, grouped into three semantic categories. This list is thought to be not only relevant and comprehensive but methodologically manageable in the disciplines of discourse analysis and corpus linguistics. An additional value of this piece of work is that the three stages can be applied to different subjects and corpora.

Keywords: Political Discourse, Corpus Linguistics, Discourse Analysis, Poverty and Social Exclusion

Anthony, L. (2014). AntConc (Version 3.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Baker, P. (2004). Querying keywords : questions of difference, frequency and sense in keywords analysis.*Journal of English Linguistics*, 32(4), 346–359.

Baker, P. et al. (2013). *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge: Cambridge University Press.

Fowler, R. (1996).On Critical Linguistics. In, Caldas-Coulthard, C. and Coulthard, M. (ed.), *Texts and Practices: Readings in Critical Discourse Analysis*. London: Routledge.

Kilgarriff, A. et al., (2004). The Sketch Engine.In Proceedings of EURALEX, *Lorient, France*.105-16.

Marchi A and Taylor C (2009) 'If on a winter's night two researchers:A challenge to assumptions of soundness of interpretation. *CADAADJournal*. 3(1): 1–20.

Stubbs, M. (1997). Whorf's children: critical comments on critical discourse analysis. In Wray, A. and Ryan, A. (eds) Evolving models of language. Clevedon: Multilingual Matters.

Van Dijk, T. (2004).Macro Context. In U. Lottgen, D. and Sánchez, J. (Eds.), *Discourse and International Relations*. Bern: Lang.

Widdowson, H. G. (2004). *Text, context, pretext*. Oxford: Blackwell.

Action and interaction through the expression of emotions in CMC: a corpus analysis of chats by English, French and Italian native speakers

This research investigates how interaction works when English, French and Italian native speakers express emotions in computer-mediated conversations. According to Oatley and Johnson-Laird (1996), "emotions are communications within the brain and among individuals" (p. 2). Furthermore, Tomasello (2008) argues that interaction is based on motivations of *requesting*, *informing*, and *sharing*. In particular, sharing means "I want you to feel something so that we can *share attitudes/feelings together*" (p. 87, italics in original). Based on the idea that surprise is the psychophysiological disruption emotions generate from (XX, 2015), this research examines the spontaneous expression of surprise. A qualitative study was conducted on a corpus of three hundred Facebook dialogues (one hundred for each language) where surprise was elicited.

By adopting a systemic approach and examining the order in which the speaker reacts to (*reaction*), comments on (*comment*) and wonders about (*question*) the surprising event, it was expected that different ways of conveying surprise exist and that their use may vary according to the medium used to communicate (Kramsch, 2009). Attention was focused on the features peculiar to CMC – e.g. smileys – and to surprise – e.g. disruption and intensity – (see Kövecses, 2003; Zammuner, 1998). More precisely, the objective was to see how these chat-language codes are employed to convey surprise. Each expression was accompanied by a short description including the speakers' level of intimacy and the sequence of the *reaction-*, *comment-* and *question-segments*.

> "Wow now I really want to go there! Did you take that picture??"
> (Intimacy: 1; Sequence: reaction-comment-question)

The objective was to examine the speaker's motivation behind the expression of this instinctive reaction and its elicitation in his/her addressee(s). Furthermore, my goal was to see whether it was possible to define a semantic pattern peculiar to the expression of surprise and computer-mediated communication, and common to the languages under investigation.

The results of this study show that CMC, where written and oral language enter in contact, gives users the possibility to modulate their productions and, as a consequence, the way they act on both their environment and addressee(s). This analysis revealed that the speakers' level of intimacy plays a crucial role in the expression of surprise: the more intimate speakers are, the more intense surprise reactions will be. Furthermore, the fact that punctuation and vowels, which reproduce the intonation of the utterance, are used to convey the intensity of reactions reveals that it is mainly through intonation that we express surprise.

References:

XX (2015). The computer-mediated expression of surprise: a corpus analysis of chats by English and Italian native speakers and Italian learners of English. *Review of Cognitive Linguistics 13*(2), 383-414.

Kövecses, Z. (2003). *Metaphor and Emotion: Language, Culture, and Body in Human Feeling*. Cambridge University Press.

Kramsch, C. J. (2009). *The Multilingual Subject: What Foreign Language Learners Say About their Experience and why it Matters*. Oxford University Press.

Oatley, K., & Johnson-Laird, P. N. (1996). The communicative theory of emotions: Empirical tests, mental models, and implications for social interaction.

Tomasello M. (2008). *Origins of Human communication*. Cambridge(MA)-London: Mit Press.

Verbless Sentences: A corpus-based contrastive study

Although the absence of the verb has fascinated linguists for ages, the difficulty of automatic processing has meant that most analyses have relied on fragmented data. This paper presents a new method of verbless sentence extraction and investigates the linguistic feature through parallel-text corpora. Profound cross-linguistic differences make it particularly relevant to compare Russian, a language that permits the most liberal use of verbless sentences among the Indo-European family, with English, known for its dependency on the verb phrase (McShane 2000). Following the contrastive analysis methods of Guillemin-Flescher (2003), this study relies on the principle that cross-linguistic comparison reveals linguistic constraints and that by studying the reoccurring choices of translators it is possible for these constraints to surface. I analyze the translated equivalents of verbless sentences in their context with the aim of getting insight into the semantics of the absence of the verb.

The results are based on a 120,000-word corpus specially created for this study. The pilot corpus includes Dostoyevsky's colloquial, dialogue-centered *The Brothers Karamazov* and two English translations. Automatic extraction was done using Trameur (Fleury & Zimina 2014) and evaluated against manual data. Multi-layered custom annotation made it possible to overcome the typical problems of fixed annotation tagsets and verb-centric syntactic modeling associated with previous studies (e.g. Landolfi et al. 2010). Data was aligned by paragraph to reveal several translation equivalents. Qualitative and quantitative analyses of the Russian-English correspondences followed. Characteristic elements of verbless/verbal sentences were also found using statistical corpus tools.

Quantitative findings confirmed the statistical significance of verbless sentence frequency differences; surprisingly, verbal ellipses were overrepresented in English.

Qualitative analysis revealed an English trend of activating contextually implied topics, particularly over distances. Example (1) illustrates a pattern in which a subject representing the topic is lexically evoked. In Russian, the omission of the extra-linguistically implied topic ('the woman') is carried over both (a) and (b); whereas in English, the topic is omitted in (b) only after having been established in the linguistic-context in (a). This trend influences the realization of the verb.

(1) (a){Immediately after the woman tells a story about herself, the elder asks her:}
Izdaleka?
from.far.away-ADV
'Have you come from far away?'

(b){Woman:}

Za	**pyatsot**	**verst**	**otseleva.**
over-PREP	five.hundred-NUM-ACC	verst-F-PL-GEN	from.here-ADV

'Over three hundred miles from here.'

Furthermore, a phenomenon of 'predication transformation' was recorded. Ellipsis of the non-copular verb *'come'* in (b) results in a syntactically verbless sentence which is, according to Hengeveld (1992), a case of semantically verbal predication. Despite the literal translation, the verbal predication in (b) was originally non-verbal in Russian.

One theoretical implication is that the verbal/non-verbal dichotomy is inadequate for a cross-linguistically stable definition of predication. Combining Lambrecht (1994), Touratier (2008) and Danon-Boileau & Morgenstern (2008), I propose to account for predication in terms of a linguistically-explicit focus and contextually-available support.

Keywords: parallel corpora, verbless sentences, predication, information structure, English, Russian

References

Danon-Boileau, Laurent & Aliyah Morgenstern. 2008. Peut-on parler de prédication dans les premiers énoncés de l'enfant? *Faits de Langues: La Prédication* 31-32, 57-65.

Fleury, Serge & Maria Zimina. 2014. Trameur: A framework for annotated text corpora exploration. In Lamia Tounsi, Rafal Rak (eds.), *Proceedings of COLING 2014 the 25th International Conference on Computational Linguistics: System Demonstrations, August 2014, Dublin, Ireland,* 57-61.

Guillemin-Flescher, Jacqueline. 2003. Théoriser la traduction. *Revue française de linguistique appliquée* 8(2), 7-18.

Hengeveld, Kees. 1992. *Non-verbal Predication: Theory, typology, diachrony*. Berlin: Mouton de Gruyter.

Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topic, focus and the mental representation of discourse referents.* Cambridge: CUP.

Landolfi Annamaria, Carmela Sammarco & Miriam Voghera. 2010. Verbless clauses in Italian, Spanish and English: a Treebank annotation. In Sergio Bolasco, Isabella Chiari, Luca Giuliano (eds.), *Proceedings of JADT 2010 the 10th International Conference on Statistical Analysis of Textual Data, June 2010, Rome, Italy,* 1187-1194.

McShane, Marjorie. 2000. Verbal ellipsis in Russian, Polish and Czech. *The Slavic and East European Journal* 44(2), 195-233.

Touratier, Christian. 2008. Que faut-il entendre par prédication et prédication seconde? *Faits de Langues: La Prédication* 31-32, 13-22.

Etude quantitative des*swearwords* dans les documentaires musicaux : démarche heuristique,questions méthodologiqueset limites théoriques

Nous proposons d'adopter une démarche réflexive concernant l'étude quantitative que nous avons menée dans le cadre de notre travail de thèse sur l'emploi des mots grossiers (*swearwords*) au sein d'un corpus composé de documentaires musicaux portant sur la musique populaire anglophone au 20ème et 21ème siècle : rock, folk, blues, punk, électro.

Il nous semble pertinent d'adopter cette démarche afin d'en saisir la portée et les limites, dont certaines ont précisément permis de mettre en lumière la nécessité d'une articulation de cette étude avec un travail qualitatif sur la notion de transgression verbale, envisagée dans le cadre de la linguistique interactionnelle et de l'analyse du discours. En effet, l'hypothèse de départ consistait à interroger le caractère supposé essentiel de la transgression au sein de la culture rock (Chastagner, 2011, 40) et de chercher à voir quel rôle l'interaction verbale, notamment médiatique, jouait dans l'établissement et la perpétuation de ce topos.

Ainsi, nous proposons de reprendre et d'exposer le cheminement méthodologique depuis les fondements théoriques de ce travail – notamment l'articulation entre les marqueurs langagiers, le profil communicatif et l'ethos (Kerbrat-Orecchioni, 2005, 304) – jusqu'aux hypothèses formulées concernant les résultats obtenus, en passant le prétraitement du corpus, le choix du logiciel (AntConc, Anthony, 2014), la constitution de la liste de formes recherchées, le traitement et la pondération des données chiffrées.

Le travail d'analyse des résultats a débouché sur un certain nombre de conclusions diversement prévisibles concernant la plasticité du tabou linguistique au sein du corpus– la grossièreté des punks étant par exemple moins surprenante que celle des Beatles. Nous avons cependant pu avancer différents critères explicatifs plus pertinents pour saisir ce phénomène : sous-genre musical dominant au sein des films, poids relatif de la musique et des images d'archives, environnements interactionnels, historiques et filmiques. Nous proposons de détailler ces résultats et leur apport à l'analyse qualitative précédemment évoquée sans perdre de vue les deux écueils dont parle Catherine Kerbrat-Orecchioni au sujet de la « recherche de

l'ethos » : la sur-généralisation caricaturale et la sous-généralisation désordonnée et anecdotique (*ibid.*, 306).

Références

ANTHONY L., 2014, AntConc (Version 3.4.3), WasedaUniversity, Tokyo, (http://www.laurenceanthony.net/)

CHASTAGNER C., 2011, *De la culture rock*, Paris, Presses universitaires de France.

KERBRAT-ORECCHIONI C., 2005, *Le discours en interaction*, Paris, A. Colin.

Corpus oraux, prosodie et linguistique pragmatique : l'exemple de « yes »

Cette étude de corpus a pour objectif de montrer à propos de « yes » :

- comment la réalisation prosodique d'un marqueur a priori peu polysémique se trouve associée à des sens en emploi – que nous nommerons interprétations-types ou emplois-types – très hétérogènes, et même parfois opposés (cas des « yes » interprétésconversationnellementcomme des « no »),
- comment l'étude de tels phénomènes, mais aussi d'une polysémie « prosodique » complètement lexicalisée (Calhoun & Schweitzer, 2012, Petit, 2009), peut être menée à partir de corpus de plusieurs centaines d'emplois de ces marqueurs en automatisant le processus de navette entre caractérisation prosodique et caractérisation sémantique et pragmatique.

Nous montrerons dans un premier temps comment une analyse prosodique des contours associés à « yes » peut être menée tant qualitativementque quantitativement à l'aide d'instruments comme PRAAT et d'un outil de stylisation automatique et d'étiquetage prosodique d'unités linguistiques.

Nous illustrerons ensuite comment l'analyse pragmatique permet d'identifier la nature des commentaires associés à ces contours, sans oublier de reconnaître la très grande complexité de ceux-ci, ni le fait qu'ils ne se contentent pas de marquer le rapport à ce qui est dit du locuteur mais traduisent l'ensemble de la relation d'interlocution, au point d'être souvent polyphonique.

Nous montrerons surtout que l'impossibilité quasi-systématique d'associer directement un motif de contour et une interprétation tient en réalité à ce que la prosodie apporte une information bien plus précise que les interprétations testées. Cela rend méthodologiquementnécessaire une navette constante entre caractérisation sémantique/pragmatique et caractérisation prosodique. Mais permet aussi d'affinerconsidérablement la caractérisation sémantico pragmatique des emplois.

Nous présenterons ensuite les méthodes de discrimination automatique des sens en emplois qui permettent de prouver la discriminabilité prosodique de ceux-ci, et décrirons la chaîne de traitement des données qui à partir de la constitution de base d'emplois et grâce à l'utilisation de techniques de classification automatique permet de tels résultats.

Enfin, nous illustrerons l'ensemble de ce protocole pour ce qui concerne la discrimination des emplois *convaincus* et *non convaincus* de « yes » (interprétation-type) mais aussi la variation interne aux emplois de conviction, le rôle des pauses précédents le « yes », et les collocations avec d'autres marqueurs.

Nous espérons ainsi, enré-interrogeant la notion de marquage, montrercomment celle-ci est complexe et primordiale.

Bibliographie

Calhoun, S & Schweitzer, A. (2012). "Can Intonation Contours be Lexicalised? Implications for Discourse Meanings". In G. Elordieta& P. Prieto (eds.) *Prosody and Meaning* (Trends in Linguistics), De Gruyter Mouton.

Hirst, D.J. (2005), "Form and function in the representation of speech prosody", *SpeechCommunication*46 , 334–347,

Lacheret-Dujour A.&Beaugendre F. (2002), *La prosodie du français*, CNRS Editions, Paris.

Mello H., Panunzi A.&Raso T. (2012). *Pragmatics and prosody: illocution, modality, attitude, information patterning and speech annotation*. Firenze University Press.

Petit, M (2009). *Discrimination prosodique et représentation du lexique : application aux emplois des connecteurs discursifs*, Thèse de doctorat, Université d'Orléans.

Vyvyan E. (2015). A unified account of polysemy within LCCM Theory, *Polysemy: Current Perspectives and Approaches,* Volume 157, April 2015, pp. 100–123

Solving long-standing semantic annotation issues with word vectors

Keywords: annotation, deep learning, neural networks, semantic classes, semantics, word vectors

Whether manual or (semi-)automatic, the semantic annotation of a comprehensive corpus or a large dataset is a hard and time-consuming task. One long-standing challenge is the resolution of lexical and syntactic ambiguities. One option favored by corpus semanticists is to use a set of predetermined classes, such as Levin (1993) for verbs or Dixon and Aikhenvald (2004) for adjectives.

Because such semantic classes are generally broad and determined *a priori*, they hardly ever match the *ad hoc*, contextual meanings of their targets. By way of example, Table 1 is a sample dataset compiled from the British National Corpus (2007). It contains a sample of adjectives that occur in *quite/rather* constructions. Each adjective is automatically annotated with the UCREL Semantic Analysis System (USAS). As expected, the contextual meanings of highly polysemous adjectives like *hot* and *cold* are poorly captured by the limited range of possibilities of the generic tagset: e.g. *hot* in *a rather hot seller* has little to do with 'Temperature'.

Cutting-edge unsupervised learning algorithms have recently offered a solution to this problem. Once trained on a very large corpus, these algorithms produce distributed representations for words in the form of vectors. Words or phrases from the vocabulary are mapped to vectors of real numbers (Table 2). Words with similar vector representations have similar meanings.

In this paper, I introduce two state-of-the-art unsupervised learning algorithms for obtaining vector representations for words: `word2vec` (Mikolov, Chen, et al. 2013; Mikolov, Sutskever, et al. 2013; Mikolov, Yih, et al. 2013) and `GloVe` (Pennington et al. 2014). Based on neural networks, these models *(a)* learn word embeddings that capture the semantics of words by incorporating both local and global corpus context, and *(b)* account for homonymy and polysemy by learning multiple embeddings per word.

I use a 840-billion-token set of word vector representations, which have been pre-trained with `GloVe`, to semantically annotate adjectives intensified by *quite* vs. *rather* in the BNC. Based on a previous case study (author, 2015), I cluster the adjectives with regard to their vector profiles. As expected, a much finer clustering is obtained when the intensified adjectives are represented as word vectors than when they are arbitrarily and artificially grouped by means of predetermined semantic classes.

Word vectors have important implications for solving semantic annotation issues. They suggest decisive opportunities for future research in corpus building, dataset compilation and, more generally, corpus-based semantics.

Table 1: a sample data frame with USAS annotation of adjectives (BNC XML)

corpus file	context	adjective	semantic tag	semantic class
K1J.xml	*quite a hot shot*	hot	O4.6+	Temperature_Hot_on_fire
G2W.xml	*a rather hot seller*	hot	O4.6+	Temperature_Hot_on_fire
KRT.xml	*a quite clear position*	clear	A7+	Likely
J0V.xml	*quite a clear understanding*	clear	A7+	Likely
CHE.xml	*quite a clear view*	clear	A7+	Likely
FEV.xml	*quite a clear picture*	clear	A7+	Likely
EWR.xml	*a quite clear line*	clear	A7+	Likely
CRK.xml	*quite a clear stand*	clear	A7+	Likely
HA7.xml	*a rather clouded issue*	clouded	O4.3	Colour_and_colour_patterns
KPV.xml	*quite a cold day*	cold	O4.6-	Temperature_Cold
G3B.xml	*a rather cold morning*	cold	B2-	Disease
AB5.xml	*a rather cold person*	cold	O4.6-	Temperature_Cold
CDB.xml	*rather a cold note*	cold	O4.6-	Temperature_Cold
K23.xml	*a rather colder winter*	colder	O4.6–	Temperature_Cold

Table 2: a sampled list of word vectors (BNC, 8 dimensions out of 50)

word	**dimension 1**	**dimension 2**	**dimension 3**	**dimension 4**	**dimension 5**	**dimension 6**	**dimension 7**	**dimension 8**	...
...	...	...	...	...	...	...	...	...	...
fears	0.006628	0.271163	0.543003	-0.272209	-0.846799	0.842019	0.542897	-0.046692	...
fearsome	0.006381	0.271721	0.077205	-0.199816	-1.073080	0.358058	0.117449	-0.243042	...
fearsomely	-0.277312	0.330456	-0.322540	0.031834	0.515634	-0.156242	0.589139	0.199485	...
feart	-0.197516	-0.060900	-0.382982	-0.067519	0.483453	-0.355956	0.175088	0.229981	...
feasant	-0.059284	0.130193	0.063192	0.154776	0.279475	-0.109160	0.404007	-0.024545	...
feasibility	0.529552	0.392431	-0.516653	0.208211	-0.589757	0.078419	-0.373397	0.592607	...
feasible	-0.204501	0.018642	-1.059382	0.256092	0.175842	0.260492	0.634680	0.128659	...
feasibly	-0.759028	-0.290727	0.136129	0.406661	0.815390	0.161068	0.426249	0.014057	...
feast	-0.236816	0.286594	1.195448	0.491160	-0.391191	0.618281	-0.400367	-0.479711	...
feasted	0.024514	-0.587153	0.176844	0.175131	0.443137	0.500882	-0.446191	0.590365	...
feasting	0.417787	-0.171134	0.610454	-0.310351	-0.865423	0.681014	-0.098815	0.043854	...
feastings	-0.249031	-0.158092	0.122720	0.086465	0.413110	-0.242642	0.003703	0.278281	...
feasts	-0.074219	0.019096	0.526855	-0.322929	-0.304485	0.459320	-0.589023	0.333732	...
feat	0.571409	0.341399	-0.167423	0.681122	-0.679442	0.291273	0.388183	-0.186563	...
feather	-0.334491	0.110777	0.263099	-0.699270	-0.057790	-0.176001	-0.023516	-0.343892	...
featherbed	-0.264394	0.597869	0.095059	0.303135	0.383182	-0.394334	0.203389	0.004839	...
feathered	0.608165	0.068559	0.217317	-0.039874	-0.353111	-0.079071	-0.300045	-0.502757	...
feathering	0.170195	-0.515068	-0.179758	0.122030	0.193888	0.533021	0.272064	0.315104	...
featherlight	-0.117239	-0.272356	-0.018808	-0.241152	0.239466	-0.290264	0.089445	0.282308	...
feathers	0.320718	-0.078353	0.198964	0.158623	-0.459843	0.166611	0.684569	-0.837097	...
...	...	...	...	...	...	...	...	...	...

References

Dixon, Robert M. W. and Alexandra Y. Aikhenvald (2004). *Adjective classes: a cross-linguistic typology*. Oxford: Oxford University Press.

Levin, Beth (1993). *English verb classes and alternations : a preliminary investigation*. Chicago ; London: University of Chicago Press.

Mikolov, Tomas, Kai Chen, et al. (2013). "Efficient Estimation of Word Representations in Vector Space." In: *CoRR* abs/1301.3781. URL: `http://arxiv.org/abs/1301.3781`.

Mikolov, Tomas, Ilya Sutskever, et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality." In: *CoRR* abs/1310.4546. URL: `http://arxiv.org/abs/1310.4546`.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic regularities in continuous space word representations." In: *Proceedings of NAACL-HLT*, pp. 746–751. URL: `http://www.aclweb.org/anthology/N/N13/N13-1090.pdf`.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1532–1543. URL: `http://www.aclweb.org/anthology/D14-1162`.

The British National Corpus (2007). *BNC XML Edition*. Version 3. URL: `http://www.natcorp.ox.ac.uk/`. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

A corpus-driven approach to native and learner spoken fluency: The contribution of pauses

Research into spoken fluency – the smooth, continuous delivery in oral production (Chambers 1997) – has traditionally focused on fine-grained qualitative analyses of a very limited set of speakers, data and/or phenomena. Recently, however, it has undergone profound changes. Thanks to the increasing number of large and representative spoken and speaking corpora (without/with access to the acoustic signal) (see Ballier and Martin 2015, 110), along with substantial improvements in speech technologies, it is now possible to combine quantitative and qualitative corpus analyses of (several aspects of) fluency. This association has yielded significant new insights into spoken behaviour (e.g. Götz 2013; Gut 2009; Osborne 2007).

This paper adopts such a combined approach and draws on two comparable spoken corpora of native and advanced learner English, namely the *Louvain Corpus of Native English Conversation* (LOCNEC, De Cock 2004) and the French component of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI, Gilquin, De Cock, and Granger 2010). Both contain 50 interviews with university students: English native speakers (LOCNEC) or foreign language learners of English (LINDSEI). The corpora (totalling c. 30 hours of recorded speech) have been time-aligned and annotated with EXMARaLDA (Schmidt and Wörner 2014) for a dozen fluency features, including filled and unfilled pauses (measured in milliseconds), restarts, repetitions and truncations (Dumont 2015).

The study first presents the overall quantitative distribution of these fluency features. It then zooms in on filled (FPs) and unfilled pauses (UPs), two features that are claimed to be particularly discriminant between fluency levels (cf. CEFR descriptors; Council of Europe 2001), and examines their qualitative use in isolation as well as in larger cluster-like sequences, e.g. a FP used conjointly with a discourse marker (*at first <u>(0.310) well</u> I was very hesitant*), or an UP within a repetition (*I think <u>we'll (0.540) we'll</u> leave together*). Based on the hypothesis that both native speakers and advanced learners may achieve varying (and potentially overlapping) degrees of fluency, native and learner data are analysed together, making no a priori presupposition as to the intrinsic benchmarking property of native speaker data.

Preliminary quantitative results indicate that among the dozen fluency features under investigation, FPs and UPs are by far the most frequent, making up together about 50% of the annotated features. However, a high rate of pausing is found in both native speakers and advanced learners, thereby shedding doubt on their use as level discriminator in the CEFR. The qualitative analysis of FPs and UPs reveals that, for all speakers (native and non-native), pauses are more frequently used in combination with (an)other fluency feature(s) than in isolation, and that, among the wide variety of possible clusters, the "bigrams" FP+UP and UP+FP are especially common.

<u>References</u>

Ballier, Nicolas, and Philippe Martin. 2015. 'Speech Annotation of Learner Corpora'. In *The Cambridge Handbook of Learner Corpus Research*, edited by Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, 107–34. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press.

Chambers, Francine. 1997. 'What Do We Mean by Fluency?' *System* 25 (4): 535–44.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. 3rd printing. Cambridge: Cambridge University Press.

De Cock, Sylvie. 2004. 'Preferred Sequences of Words in NS and NNS Speech'. *Belgian Journal of English Language and Literature (BELL)* 2: 225–46.

Dumont, Amandine. 2015. 'Designing and Implementing a Multilayer Annotation System for (dis)fluency Features in Learner and Native Corpora'. Paper presented at the Corpus Linguistics Conference, Lancaster (UK).

Gilquin, Gaëtanelle, Sylvie De Cock, and Sylviane Granger, eds. 2010. *LINDSEI. Louvain International Database of Spoken English Interlanguage*. Presses Universitaires de Louvain. Louvain-la-Neuve.

Götz, Sandra. 2013. *Fluency in Native and Nonnative English Speech*. Studies in Corpus Linguistics (SCL) 53. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Gut, Ulrike. 2009. *Non-Native Speech: A Corpus-Based Analysis of Phonological and Phonetic Properties of L2 English and German*. Edited by Thomas Kohnen and Joybrato Mukherjee. English Corpus Linguistics 9. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

Osborne, John. 2007. 'Investigating L2 Fluency through Oral Learner Corpora'. In *Spoken Corpora in Applied Linguistics*, edited by Mari Carmen Campoy and María José Luzón, 57:181–97. Linguistic Insights. Studies in Language and Communication. Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien: Peter Lang.

Schmidt, Thomas, and Kai Wörner. 2014. 'EXMARaLDA'. In *The Oxford Handbook of Corpus Phonology*, edited by Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 402–19. Oxford University Press.

According to Bhatia (2005), the phenomena of vagueness in human thought and communication are traditionally considered as an object of philosophical interest, but, being "a contrastive property of natural language", vagueness draws considerable attention from scholars working in different disciplines. Following previous studies, such as Di Carlo (2013), Warren (1988), and Bhatia (2005), the present study investigates vagueness-related phenomena through Russian-English parallel and comparable corpora from a linguistic perspective.

The use of ambiguous words in international communication has become a stable tendency, which aroused my interest in this research area and led me to pose the following questions: Why does vagueness play such an important role in contemporary international communication? What do we understand by vagueness? What are the components included in the meaning of vagueness? How do official texts and documents bear responsibility by implying vagueness? By what means does vagueness appear in documents: through modal verbs - mainly, the case of *shall* (see Example 1, where *shall* is used as a tool for conveying obligation, and regulating behavior, however the use of *shall* is contextually depersonalized, which makes the utterance imprecise and creates ambiguity) and *should* will be considered, through weasel words and adjectives (see Example 2, which shows the use of some weasel words, implying a flexible meaning and subjective interpretation, depending on how a reader will define what is *necessary*, *crucial* or *available* in the suggested context), or any other categories?

(1) *…The diplomatic courier, who **shall** be provided with an official document indicating his status and the number of packages constituting the diplomatic bag, **shall** be protected by the receiving State in the performance of his functions. He **shall** enjoy person inviolability and **shall** not be liable to any form of arrest or detention…* (Vienna Convention 1961)

(2) *…with law reform in the field of commercial law, also takes note that the Commission urged the Secretary-General to take steps to ensure that the comparatively small amount of additional resources **necessary** to meet a demand so **crucial** to development are made **available** promptly, and recalls paragraph 48 of its resolution 66/246 of 24 December 2011 regarding the rotation scheme of meetings between Vienna and New York…*(A/RES/67/88)

The research question is to find out to which extent these categories and instruments can be considered vague or indeterminate in terms of their actual usage in diplomatic and normative texts. The meaning and use of the modals, weasel words, adjectives and adverbs are analyzed and compared on the basis of semantic and pragmatic parameters. Additionally statistical analysis is implemented, allowing us to understand the quantitative structure of the data in terms of frequency and contingency. Data for this research is mainly drawn from two types of Russian-English corpora: Parallel and Comparable. Parallel corpora include the United Nations resolutions from the Sixth Committee (resolutions of the 67^{th}, 68^{th}, 69^{th} Sessions), which stand for the consideration of legal questions in the General Assembly. Comparable corpora are supplied by the collected documents from different types of discourse: legal, diplomatic, normative discourse. They contain diplomatic conventions, communiqués, letters of credence, and diplomatic notes.

Methods used for this study include corpus linguistics methods and methods of annotation (part-of-speech tagging). Annotated corpora will be processed using characteristic elements computation, and factorial analysis. Analysis of the chosen data is carried out by using such tools as AntConc, TreeTagger, Le Trameur, *R*-Studio. However, it seems relevant to conduct this research on different annotation levels, including, for instance, morphosyntactic tagging.

Following the requirements of the conference "New Approaches to Corpus in English Linguistics", the presentation will include English and Russian data drawn from the corpora used for the present study.

Keywords: *legal language, diplomatic language, parallel corpus, comparable corpus, normative texts, vagueness, modal verbs, weasel words.*

REFERENCES:

Bhatia, V. K. (2005). *Vagueness in normative texts* (Vol. 23). Peter Lang.

Deschamps, K., & Smessaert, H. (2009). The Logical-Semantic Structure of Legislative Sentences. *Comparative Legilinguistics. International Journal for Legal Communication*, (1), 73-87.

Di Carlo, G. S. (2013). *Vagueness as a political strategy: Weasel words in security council resolutions relating to the Second Gulf War*. Cambridge Scholars Publishing.

Fleury, S., & Zimina, M. (2014, July). Trameur: A Framework for Annotated Text Corpora Exploration. In COLING (Demos) (pp. 57-61).

Gries, S. T. (2009). What is corpus linguistics?. *Language and linguistics compass*, 3(5), 1225-1241.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Palmer, F. R. (1990). *Modality and the English modals*. Longman.

Warren, B. (1988). Ambiguity and vagueness in adjectives. *Studia linguistica*,*42*(2), 122-172.

Corpus Data cited in the abstract proposal:

http://research.un.org/en/docs/ga/quick/regular/67

http://legal.un.org/ilc/texts/instruments/english/conventions/9_1_1961.pdf

Constitution et exploitation de corpus oraux : illustration

Cette communication se penche sur les corpus oraux, leur spécificité, leur constitution et leur exploitation, en l'illustrant à l'aide d'exemples.
Un corpus linguistique est une collection de données organisées selon des critères linguistiques bien précis et pouvant servir d'échantillon de langage (cf. Habert et al. 1997 ; Sinclair 1996). Le corpus doit pouvoir être également réutilisable (Gibbon, 1997). Le terme de « corpus oral » sera pris ici dans un sens large : tout corpus d'enregistrements de productions langagières (corpus oral à proprement parler, corpus de parole, corpus phonologique, corpus dialectologique…), par opposition aux corpus écrits.
La constitution d'un corpus (oral) nécessite plusieurs phases : a) un travail préparatoire de réflexion : but, méthodologie, métadonnées/documentation, récupérabilité (mise à disposition)…b) une collecte des données elles-mêmes (enregistrements audio / vidéo) c) une documentation, des transcriptions et annotations d) une curation (archivage, accessibilité). Les données enregistrées peuvent être de différents type : conversations spontanées, semi-spontanées (émissions, cours), spontané contrôlé (map task), lecture de textes, de mots, logatomes… La documentation inclut le but, la méthodologie adoptée, des données sur les locuteurs, sur la situation d'interlocution, etc. Les transcriptions et annotations (manuelles ou effectuées à l'aide de logiciels comme SPASS ou PRAAT), le plus souvent alignées sur le signal, peuvent être de diverses natures : transcription orthographique (enrichie ou non) – qui pose le problème de la représentation écrite de l'oralité (Blanche-Benvéniste & Jeanjean, 1987) ; signal sonore (transcription phonétique/phonologique, indications prosodiques), éléments linguistiques (tours de parole, chevauchements…), éléments non linguistiques (communicatifs ou non : toux, rires…), éléments non sonores (gestuelle), ponctuation ou non, étiquetage morphologique et syntaxique… Nous illustrerons la constitution d'un corpus oral à l'aide du corpus PAC, corpus en cours de développement, qui vise l'étude des variétés des systèmes phonologiques anglais (segmentaux ou suprasegmentaux). Nous en détaillerons la méthodologie.
Le corpus peut être exploité à différentes fins : le corpus de validation sert à vérifier une hypothèse qui existe indépendamment tandis qu'avec le corpus heuristique, le linguiste « part à la rencontre de l'inconnu avec un questionnement, mais pas de solution » (Scheer, 2004). Nous illustrerons l'emploi heuristique d'un corpus oral à travers une étude sur un corpus de référence de grande taille (ICE-GB), et l'emploi comme corpus de validation par le biais d'une étude sur le corpus PAC. La première explore les interactions entre syntaxe et prosodie dans le cadre des structures clivées ; la deuxième concerne le système vocalique du Donegal.

Références :

Blanche-Benvéniste & Jeanjean, 1987, *Le français parlé : transcription et éditions,* Paris : Didier Erudition.

Gibbon, D., Moore, R. & Winski, R., (eds.), 1997, *Handbook of Standards and Resources for Spoken Language Systems, vol. 1 : Spoken Language Systems and Corpus Design*, Berlin : Mouton De Gruyter.

Habert, B., Nazarenko, A. & Salem, A., 1997, *Les linguistiques de corpus*. Collection U, série « Linguistique », Paris : Armand Colin.

Sinclair, 1996. *Preliminary recommandations on corpus typology,* Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).

Scheer, 2004. Le corpus Heuristique : un outil qui montre mais ne démontre pas, *Corpus 3 : Usage des corpus en phonologie*, 153-192. Mis en ligne le 02 décembre 2005, consulté le 17 janvier 2013. URL : http://corpus.revues.org/210.

Analyse de la couverture médiatique des primaires américaines par ses quotidiens nationaux : une approche lexicométrique

L'élection présidentielle américaine de 2016 constitue un évènement politique et médiatique majeur, dont la course à l'investiture démocrate et républicaine est une étape décisive. Le discours médiatique couvrant les primaires américaines est déterminant, car il permet à une large audience de se forger une opinion sur les différents projets politiques et les personnalités qui les portent.
En s'appuyant sur la théorie du cadrage médiatique ("*media frames*", Gitlin 1980, Entman 1993), notre objectif est de caractériser les schémas d'interprétation proposés par les médias dans leur traitement des primaires, c'est-à-dire leur mise en lumière des candidats et du contenu politique de leurs discours. Nous questionnerons alors la compétition des cadrages (*framing contest*, Gamson 1992) s'exerçant entre les différents candidats dans la définition des sujets importants, de leurs facteurs explicatifs et des éléments de raisonnement associés. Nous analyserons plus particulièrement la prise en charge de cette compétition par les médias américains dans leur reconstruction des thématiques soulevées, au sein d'un discours médiatique diffusé auprès des (é)lecteurs.
Sur le plan opérationnel, nous proposons d'analyser un corpus composé de plusieurs milliers d'articles portant sur ces primaires et publiés par les 10 principaux quotidiens nationaux américains, entre le 25 janvier et le 25 mars 2016. Ce corpus médiatique substantiel sera analysé à l'aide de différents outils relevant de la statistique lexicale (Lebart & Salem, 1994), notamment grâce au logiciel Iramuteq (Ratinaud & Dejean, 2009), permettant une classification hiérarchique descendante du corpus (Reinert, 1983). Cette analyse est construite sur l'identification de réseaux étroits de cooccurrences lexicales, lesquelles permettent d'approcher une dimension thématique et sémantique des discours (Mayaffre, 2015) sur la base d'indicateurs statistiques.
Il s'agit ainsi d'interroger certaines spécificités du discours médiatique américain en période électorale. Analyses contrastives, factorielles et classificatoires nous permettront de mesurer la présence d'un discours adossé aux résultats des sondages d'opinion, primaires et autres caucus (dit "horse race framing", Iyengar 1991) et de préciser les modalités discursives de la mise en débat des positions idéologiques des candidats. Cette médiation journalistique peut s'avérer biaisée (Entman 2010), en fonction de la période, de l'identité du candidat, de son parti, et de leur résonance avec celle des différents journaux, dont le contrat de communication (Charaudeau, 1997) est fondé sur le partage avec les lecteurs de certaines valeurs, y compris politiques.

Charaudeau, P. (1997). Le discours d'information médiatique : la construction du miroir social. Paris : Nathan, INA.

Entman, RM. (1993). Framing: Toward Clarification of a Fractured Paradigm. Journal of Communication, 43 (4), p. 51-58.

Entman, RM. (2010). Media framing biases and political power: explaining slant in news of campaign 2008. *Journalism*, 11(4), p. 389–408.

Gamson, W. (1992). Talking Politics. Cambridge University Press.

Gitlin, T. (1980). The Whole World Is Watching. Berkeley, Los Angeles: University of California Press.

Iyengar, S. (1991). Is anyone responsible ? How television frames political issues. The University of Chicago Press. 206 p.

Lebart, L. & Salem, A. (1994). Statistique textuelle. Paris : Dunod.

Mayaffre, D. (2014). « Plaidoyer en faveur de l'analyse de données co(n)textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) », JADT 2014, Proceedings of the 12th International Conference on Textual Data Statistical Analysis, E. Née, M. Valette, J.-M. Daube, S. Fleury éd., Paris, Inalco - Sorbonne nouvelle, p. 15-32, en ligne : http://lexicometrica.univ-paris3.fr/jadt/jadt2014/01-ACTES/01-JADT2014.pdf.

Ratinaud, P., & Dejean, S. (2009). IRaMuTeQ: implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. Presented at the *Modélisation Appliquée aux Sciences Humaines et Sociales* (MASHS2009), Toulouse, France.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, VIII, (2), 187-198.

Affective Sentiment in the Non-Adoption of Transmedia Texts: A Corpus Based Investigation of Gender Difference

Keywords: Transmedia; Technology; Non-adoption; Affective barriers; Gender difference.

Jamie McKeown
The Hong Kong Polytechnic University
Jamie.mckeown@gmail.com

Fuelled by the rise of the ubiquitous web and the proliferation of digital technology, transmedia storytelling has radically altered the creative landscape. Defined as a process in which integral parts of a fiction are dispersed across multiple channels (1), transmedia offers infinite creative possibilities for a new generation of content makers, and novel modes of participation for audiences. Unfortunately, the field has been plagued with many examples of creative projects, aimed at mass participation, that have failed to gain significant traction beyond fan communities. In order to remain commercially viable, transmedia activity cannot remain the chicane of fandom. The current paper, treats the failure of the field to achieve significant participation as an issue of non-adoption (i.e. from a technological perspective), and specifically investigates affective reasons for non-adoption of digital extensions of terrestrial television shows.

Scholarly investigation of technology adoption has traditionally tended to favour notions of rational cognition (2; 3; 4) over affective considerations. In recognition of findings that claim decision making is influenced by feelings as much as rational discernment (5; 6) recent studies have begun to explore emotion as a motivational force (7; 8). Such studies, to date, have generally focused on one emotional dimension (9) i.e. hedonic motivation (fun), as a driver of adoption/non-adoption of technology.

The present paper will create a corpus of free text responses (n = 616[1]) to the following question: *Please explain why you do not watch online extensions of your favourite tv dramas?* Gender will be used as a contrastive variable. Wmatrix will be used to extract responses that fall within the semantic domain of emotion. These responses will then be manually classified according to the drivers of adoption/rejection in the Consumer Acceptance of Technology model, CAT(10), in order to identify the most evocative drivers. Finally, the entire corpus will be manually coded in order to discern the usefulness of corpus tools (i.e. Wmatrix) versus manual annotation, in capturing attitude and emotion as present in the data. In terms of substantive research questions, the present paper aims to:

1) identify emotional responses (if any) other than hedonic motivation in the non-adoption of transmedia texts;

2) map affective response according to the 6 drivers of acceptance/rejection of new technology (according to CAT);

3) identify any gender differences in relation to the previous two objectives;

[1] Online survey conducted amongst general UK population, 2012.

Finally, from a methodological perspective, the present paper attempts to:

4) explore the usefulness of corpus tools versus manual annotation in the identification of emotion and attitude in free text data.

References:

(1) Jenkins, H. (2008). *Convergence Culture*. New York: New York University Press.

(2) Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading,MA: Addison-Wesley Publishing Company.

(3) Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35, 982–1004.

(3) Rogers, E. M. (2003). *Diffusion of innovations*, 5th ed. New York: Free Press.

(5) Holbrook, M. B., & Hirschman, E. C. (1982). The experiential aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of Consumer Research*, 9, 132–140.

(6) Hartman, J. B., Shim, S. Barber, B., & O'Brien, M. (2006). Adolescents' utilitarian and hedonic web-consumption behavior: Hierarchical influence of personal values and innovativeness. *Psychology & Marketing*, 23, 813–839.

(7) Bruner II, G. C., & Kumar, A. (2005). Applying T.A.M. to consumer usage of handheld Internet devices. *Journal of Business Research,* 58, 553–558.

(8) Childers, T. L., Carr, C. L., Peck, J., & Carson, S. (2001). Hedonic and utilitarian motivations for online retail shopping behavior. *Journal of Retailing*, 77, 511–536.

(9) Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), pp. 425-478.

(10) Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Identifying speech acts in a corpus of historical migrant correspondence

A full account of the pragmatics of personal correspondence requires speech act annotation, whereby each utterance in the corpus is assigned to a category such as request, commitment, or expression of feeling. This enables the user to analyse the data in an inclusive manner, considering all examples of a given speech act regardless of how they are expressed. For large datasets it would be extremely difficult – as well as time-consuming and costly – to carry out the annotation manually. We propose to use instead an automated speech act tagger developed by De Felice (De Felice et al. 2013). The speech act tagger was originally designed for use with business emails; however the latest iteration of the tagger can be applied to other datasets – such as personal correspondence – providing a useful resource for the corpus linguistics community.

As part of an AHRC research networking project, 'Digitising experiences of migration: the development of interconnected letter collections', the speech act tagger was tested on a collection ofletters written by Irish emigrants at the end of the nineteenth century. This paper will report on the results of this trial study, demonstrating how the tagger can perform with some success even on corpora with very different characteristics. Although the dataset used for this trial study is small (just one letter series containing 99 texts), the findings show the potential for carrying out this type of analysis across larger digital archives allowing for different datasets to be compared, taking into consideration sociobiographic variables such as the author's sex, class and role within the notional familial hierarchy.

A corpus-based approach to emotions and pseudonimity in Computer Mediated Communication

In recent years, researchers have started to work on new kinds of communications, namely Computer Meditated Communication and research groups have been created dedicated to their study (e.g. AGORA, CoMeRe…). The format of computer mediated communication corpora is different from usual texts and news ways of analyzing data emerged since a number of new tools are made available. The very first studies on these corpora tended to be qualitative, based on samples, with a view to understanding these new supports and the new elements that were added to text structure; for example smileys and abbreviations had to be taken into account. The latter have been found to be really important especially in the study of emotions in Computer Mediated Communication, since they are actually supposed to convey something that can be related to emotions, which is not only descriptive but also expressive. On the other hand, studies both in French and in English were carried out on identity on the Internet, on identity abuses and how cyber criminality can be fought with language analysis (e.g. Tim Grant 2006 & 2007, Haya Bechar-Israeli 2006, François Perea 2010). But once again, these are specific cases from which qualitatively based general features and processes emerged. In the end, emotions and pseudonimity have mostly been studied separately, and there is a need to bring them together and add a quantitative dimension to the whole in order to conduct a more accurate and complete analysis.

The first section of the present proposes a corpus based approach to the expression of emotions. While it is fairly easy to quantify lexemes that describe emotions, ways to express emotions in language are really numerous and it seems difficult to adopt a quantitative approach to their various realizations. But new corpora of Computer Mediated Communication include tools that allow counting (e.g. smileys, abbreviations and repetition of letters) without necessarily requiring a context, though context may influence their interpretation. That is why the study of their collocations may be relevant to differentiate and quantify their uses in context, and thus classify their different uses related to different emotions.

The second section of this paper deals with the real world context of Computer Mediated Communication: people do not see or hear each other when communicating anymore. This brings new major issues in relation to Computer Mediated Communication, namely pseudonimity and expectation. Indeed, when classifying the quantitative data gathered for expression of emotions, the dimension of the unknown has to be considered. Both participants surely have expectations that will influence the discourse they produce since they might or might not know each other. By comparing the data of people already sharing their identity and tastes on different subjects, with those of people discovering the addressee behind their pseudonym and their written language, we propose to verify our hypothesis on the role of pseudonimity in expressing emotions on Computer Mediated Communication.

Finally, attention will be drawn on the differences between French and English languages in Computer Mediated Communication. The contrastive analysis of the data previously gathered allows us to compare how the expression of emotions can differ or look alike in both languages. Hypotheses will be proposed to explain these phenomena and future works may be carried out to verify their relevance.

References:

- Perakyla, A., & Sorjonen, M. L. (2012). *Emotion in interaction*. Oxford University Press.

- Caffi, C., & Janney, R. W. (1994). Toward a pragmatics of emotive communication. *Journal of pragmatics*, *22*(3), 325-373.

- Miceli, M., & Castelfranchi, C. (2014). *Expectancy and emotion*. OUP Oxford.

- Harris, R. B., & Paradice, D. (2007). An investigation of the computer-mediated communication of emotions. *Journal of Applied Sciences Research*, *3*(12), 2081-2090.

- Bechar-Israeli, H. (1995). FROM〈Bonehead〉 TO〈cLoNehEAd〉: NICKNAMES, PLAY, AND IDENTITY ON INTERNET RELAY CHAT1. *Journal of Computer-Mediated Communication*, *1*(2), 0-0.

- Georges, F. (2009). Représentation de soi et identité numérique. *Réseaux*, (2), 165-193.

- Perea, F. (2010). L'identité numérique: de la cité à l'écran. Quelques aspects de la représentation de soi dans l'espace numérique. *Les Enjeux de l'information et de la communication*, *2010*(1), 144-159.

- Paveau, M. A. (2012). Linguistique et numérique 4. Les écritures de Protée: identités pseudonymes.

- Woodhams, J., & Grant, T. (2006). Developing a categorization system for rapists' speech. *Psychology, Crime & Law*, *12*(3), 245-260.

- Sheridan, L. P., & Grant, T. (2007). Is cyberstalking different?. *Psychology, Crime & Law*, *13*(6), 627-640.

Title:
Toward EFL Learner Annotated Corpora: A New Markup Convention

Abstract:

The aim of this study is to provide a novel perspective for a corpus annotation scheme supported by a new markup convention specifically designed for EFL reading materials in a Japanese university. A newly developed e-learning system named iBELLEs (interactive Blended English Language Learning Enhancement system) plays a crucial role in the study.

iBELLEs, which is equipped with an interactive communication facility, enables EFL teachers to create self-defined, pedagogically significant tags instead of offering ordinary tags like POS tags, structural tags or semantic tags. Following the teacher-defined tag framework, the EFL learners are prompted to give individual, even heuristic markups to a particular reading material shared by all the participants in a face-to-face classroom session. Through the iBELLEs screen, the teacher can spontaneously observe overlapping markupsthat were given to the same text, and dynamically select the most effective teaching style during a single session. The students' markup information would show the teacher how students actually read the target text, e.g. words or sentences upon which they stumbled or which they assumed were keywords or topic sentences of the text.

What is important in the markup process using iBELLEs is the fact that a common reading material is annotated by a number of individual EFL learners in various ways.In other wordsiBELLEs is not used to create traditional "annotated learner" corpora, but completely new "learner annotated" corpora, in whichthe students' markups for an identical text may occasionally overlap or contradict each other. The students' markupswith versatile information stored in the server are used as valuable resource for the detailed pedagogical investigation with which each of theiractual reading processes is categorised in order to design more efficient teaching/learning plans.

The student's markups are used to explore the correlation betweentheir reading barriers and optimal practice style. For example, when a student's markup shows that he/she has a greater difficulty in understanding the structural properties of the target passage rather than vocabulary, he/she is encouraged to read other materials paying more careful attention to structural clues such as discourse markers or subordinate conjunctions. On the other hand, when a "learner annotated" corpus indicates that he/she has difficulty in finding the gist of a given text, the student is instructed to acquire skills to pick up topics or keywords in the target text.

This paper discusses the present state of the study as well as its further possibilities.

The case for environmental justice: tracking variation in do-it-yourself vs. web corpora

"Over the last three decades, a substantial environmental justice movement has emerged, although not always under that name" (UNEP, 2007:314). Official recognition of concerns over environmental justice (EJ) is even more recent, and integrating EJ considerations in discourse about climate change still seems particularly challenging. The aim of this talk is to show how observed variation may shed light on official discourse about climate change. Our approach to variation combines insights from corpus pragmatics –i.e. horizontal reading of a relatively homogeneous set of texts- with the more vertical methodology of corpus linguistics (Rühlemann and Aijmer, 2014). However, quantitative exploration is a first step in our study: we use basic corpus linguistics tools to track variation and outline pragmatic features which are then characterized and singled out for qualitative analyses.

We start from two sets of reports collected from the following United Nation agencies or programmes –REDD, IPCC, UNEP. Both are roughly two-million word corpora, corresponding to texts written before or after COP15 for the first set, and around COP21 for the second. Using the concordancer Antconc (Anthony, L. 2014) we look for keywords which are then used as seeds in BootCaT (Baroni, S. and Bernardini, S., 2004) in order to bootstrap two distinct corpora from the web, matching our two sets of reports. The resulting corpora are compared with a view to enhancing specific features of the UN reports.

Two series of analyses are conducted in Antconc. We first look at the use of first-person pronouns as possible pointers to performatives. Although Austin's explicit performatives are typically self-referential (Austin, 1975:32), first-person pronouns are not a straightforward feature of performatives. Thus, more qualitative analyses are needed at this step: identification-in-context (Garcia and Drescher, 2006) is performed on hundreds of concordances in order to assess their pragmatic value and distinguish true performatives from other, more or less conventional uses. Interestingly enough, while there is some variety in our web corpora, performatives are almost exclusively found in the conclusions and acknowledgements of UN reports. The second series of analyses focuses on other narrative features of discourse about climate change as evidence for varying degrees of reliance on existing or ad-hoc scenarios (see e.g. Obasi and Töpfer, 2000; Maranville et al., 2009).

References

Anthony, L. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Austin, J.L. 1975. *How to Do Things with Words*. Harvard University Press.

Baroni, M. & Bernardini, S. 2004. « BootCaT: Bootstrapping Corpora and Terms from the Web. » In *LREC*. http://www.cs.utah.edu/nlp/readinglist/BaroniB04.pdf.

Garcia, P. & Dresher, N. 2006. Corpus-Based Analysis of Pragmatic Meaning, in *Corpus Linguistics: Applications for the Study of English*. Peter Lang.

Maranville, Angela R., Tih-Fen Ting, et Yang Zhang. 2009. « An environmental justice analysis: superfund sites and surrounding communities in Illinois ». *Environmental Justice* 2 (2): 49‑58.

Obasi, G.O.P, & Töpfer. 2000. *Emissions Scenarios: Summary for Policymakers : A Special Report of IPCC Working Group III*. [Geneva]: WMO (World Meteorological Organization) : UNEP (United Nations Environment Programme).

Rühlemann, C. & Aijmer, K. 2014. Introduction. Corpus pragmatics: laying the foundations, in *Corpus Pragmatics. A Handbook*. Aijmer, K. and Rühlemann, C. (eds.), Cambridge University Press.

United Nations Environment Programme (UNEP), éd. 2007. *Global environment outlook: environment for development, GEO 4*. Nairobi, Kenya : London: United Nations Environment Programme ; Stationery Office [distributor].

Do auxiliaire + prédicats d'action : une propension nulle en anglais contemporain

La présente étude est consacrée au verbe *do* dans ses emplois auxiliaires en anglais contemporain et à sa possible collocation avec les prédicats d'action. Dès le XIIIe siècle, les occurrences du *do* périphrastique (ancêtre de l'auxiliaire) suivi d'un infinitif révélaient une propension du verbe à apparaître avec des prédicats désignant des procès dynamiques, propension vraisemblablement liée à son sens d'accomplissement (*achieve*, *perform*, cf. [1]).

> [1] (?c1450) Wright, T. (ed.) (1906) *The Book of the Knight of La Tour-Landry* rev. edn (EETS, Ordinary series 33), 2–24 (emprunté à Visser 1963-1973)
>
> *And so thei dede bothe deseiue ladies and gentilwomen, and bere forthe diuerse langages on hem.*
>
> and so they did both deceive ladies and gentlewomen and make diverse allegations about them

D'après Denison (1985 : §4.4.2), la majorité des prédicats associés au *do* périphrastique aux XIIIe et XIVe siècles dénotaient des procès de type action, avec environ 2 % de prédicats référant à des procès statiques uniquement. Nous nous sommes demandé si cette cooccurrence privilégiée de *do* avec des prédicats dénotant des procès dynamiques, très majoritairement téliques, se vérifiait en anglais contemporain ou si le phénomène constaté pour le *do* périphrastique s'était atténué, voire dissipé, avec la régulation du verbe en auxiliaire.

Fondant notre travail sur une analyse d'occurrences tirées du *British National Corpus* (*BNC*), nous montrerons que l'auxiliaire *do*, opérateur de l'anglais contemporain, est éminemment compatible avec les prédicats d'état, ne marquant plus l'accomplissement d'un procès nécessairement dynamique, souvent borné à droite, mais la validité d'une relation prédicative mettant en jeu tout type de prédicat. Ainsi, le *do* de *I do not!* en [2] coderait la validité de la relation prédicative <*I – want her to wear [my] suit*> en t_o, une validité niée par *not*. Devenu marqueur d'adéquation de la relation prédicative à l'extralinguistique, l'auxiliaire *do* peut aujourd'hui se combiner avec des prédicats référant à des procès aussi bien statiques que dynamiques, les reprendre, ou permettre une emphase sur la validité de leur lien avec le sujet au moment repère.

> [2] *"You want rid of her, yet you don't want her to wear your suit." "No I do not! She might taint it psychically. Interfere with the protection runes." Grimm chortled.* (*BNC*)

Nous arguerons que ce signifié d'adéquation de la relation prédicative au monde – à un moment repère déterminé par le temps porté par l'auxiliaire – participe d'un invariant sémantique d'accomplissement caractéristique du verbe *do* dans ses emplois lexicaux et auxiliaires en anglais contemporain, tout comme dans ses emplois périphrastiques dès les états anciens de l'anglais (cf. Visser 1963-1973 : §§1412-14).

REFERENCES

DAVIES, M. 2004-. *BYU-BNC: The British National Corpus*. http://corpus.byu.edu/bnc/

DENISON, D. 1985. « The origins of periphrastic DO: Ellegård and Visser reconsidered ». In R. Eaton, O. Fischer, W. Koopman et F. van der Leek (eds), *Papers from the 4th International Conference on English Historical Linguistics: Amsterdam, 10-13 April 1985*. Amsterdam & Philadelphia : John Benjamins, 45-60.

VISSER, F. T. 1963-73. *An Historical Syntax of the English Language*, 3 vol. Leiden : E. J. Brill.

Title: The interrelation between the modal *can* and the grammatical aspect of the main verb in contemporary American English – an empirical contribution to modality-aspect interaction studies.

The proposed paper presents elements of empirical, corpus-based research, in which the author traces the paradigm of modality-aspect interaction (eg. Abraham 2008, Abraham and Leiss (eds.) 2008). In order to serve this purpose, the present study investigates the interaction of the modal auxiliary *can* with the category of aspect, viewed from the grammatical perspective (Comrie 1976/2001, Binnick 1991, Higginbotham 2009). The studied linguistic phenomenon was approached with the use of the material from the *Corpus of Contemporary American English*, published by the Brigham Young University and available at *http://corpus.byu.edu/coca*.

The study deals with three potential areas of interaction between *can* and the grammatical aspect of the main verbs. For one thing, the research focuses on the modal readings of the analyzed language samples. In addition to this, the study categorizes these findings into the typological dichotomy of modalities between root and epistemic readings (Sweetser 1982, Kratzer 1991). However, all this is achieved from a prior scrutiny of the aforesaid interaction within the semantic field of modality (Kratzer 1991). The variations of the domains within this model demonstrate how the process of modality-aspect interaction is performed with reference to the modal auxiliary *can*.

One of the most significant findings of this research is the correlation between the grammatical aspect of the main verb and the type of modality expressed in the matrix predicate, which entails a variation in the interpretation of the modal reading of one modal auxiliary. With the studied language samples, the simple aspect was revealed to trigger circumstantial modal readings of *can*, thus ROOT modality. Then, the analyzed matrix predicates with *can* followed by the progressive aspect of the main verb were found to express circumstantial modal readings, thus ROOT modality. Finally, matrix predicates with the modal *can* followed by the perfect aspect of the main verb yield epistemic modal readings, thus EPISTEMIC modality.

Overall, the discussed empirical analysis proves that modality tends to interact with aspect (cf. Abraham 2008). Moreover, the findings mark out certain patterns of interaction which rely on the grammatical constructions. What is more, one may propose that further corpus-based research, concentrating on other modals and other types of aspect (eg. viewpoint or lexical aspects), may provide more findings within the realms of modality-aspect interfaces.

Key words: modality, grammatical aspect, modality-aspect interaction, American English, corpus

References:

Abraham, W. (2008) “On the logic of generalizations about cross-linguistic aspect-modality links.” [in:] W. Abraham and E. Leiss (eds.), pp. 3–13.

Abraham, W. and E. Leiss (2008) *Modality–aspect interfaces: implications and typological solutions*. - Amsterdam, Philadelphia: John Benjamins.

Binnick, R. I. (1991) *Time and the verb: A guide to tense and aspect*. New York: Oxford University Press.

Comrie, B. (1976/2001). *Aspect: An introduction to the study of verbal aspect and related problems*. Cambridge; New York: Cambridge University Press.

Higginbotham, J. (2009) *Tense, aspect and indexicality*. Oxford: Oxford University Press.

Kratzer, A. (1991): “Modality.” [in:] A. von Stechow and D. Wunderlich (eds.) *Semantics: an international handbook of contemporary research*. - Berlin: de Gruyter, pp. 639-650.

Sweetser, E. E. (1982) “Root and epistemic modals: causality in two worlds”. [in:] *Proceedings of the Eighth Annual Meeting of the Berkeley Linguistics Society* (1982). Berkley, C.A.: Berkley Linguistics Society, pp. 484-507.

The Dialogue Annotation and Research Tool (DART; Weisser 2016) is a semi-automatic annotation and analysis tool, designed to facilitate large-scale corpus-based research into pragmatics. It allows researchers to compile, annotate, and explore dialogue data, based on an annotation scheme that has evolved over a number of years. This scheme is designed to be generically applicable to as many types of spoken interaction as possible, and has already been successfully applied in the annotation of a number of corpora from various domains (see Leech & Weisser, 2003;Weisser 2010, 2014). This presentation will provide a detailed description and discussion of the annotation scheme, including its application and appliability to various types of spoken corpora.

I will begin by describing the format of the scheme, a simple and eminently readable form of XML that facilitates 'interaction' with the data. It comprises annotations on the levels of syntax, semantics, semantico-pragmatics (Searle's 'IFIDS'; cf. Searle 1969: 30), pragmatics (speech-acts), surface-polarity, a limited amount of prosodic features, as well as other interactive features such as backchanneling or overlap (see Leech & Weisser 2013 for an earlier version). This description will include a brief discussion of the scheme's genesis and development over the years, and an illustration of its advantages in comparison to other annotation formats/schemes that have been applied to the representation and annotation of dialogue data in the past (cf. Weisser 2014).

The next section will illustrate how the scheme can easily be applied in the large-scale annotation of dialogue corpora. This will include issues such as conversion from other formats and suitable pre-processing of the data to split it into appropriate units for analysis, as well as how the annotation and post-processing can efficiently be conducted within the DART environment. Some of the data used for illustration here will be drawn from task-oriented corpora, such as the SRI Amex corpus, the Switchboard corpus for unconstrained general dialogue, as well as data from some of the ICE corpora, reflecting recent attempts on my part to apply the scheme to non-native and learner Englishes.

In the final part, I will illustrate the appliability of the scheme to pragmatics- and interaction-related research by showing how the annotations can be exploited in DART in order to create speaker profiles that allow the researcher to investigate features of (in)directness & politeness, initiative, etc. (cf. Weisser 2016 forthcoming).

References:

Leech, Geoffrey, & Weisser, Martin. (2003). Generic Speech Act Annotation for Task-Oriented Dialogue. In Archer/Rayson/Wilson/McEnery (Eds.) *Proceedings of the Corpus Linguistics (2003 Conference*. Lancaster University: UCREL Technical Papers, vol. 16.

Leech, Geoffrey & Weisser, Martin. (2013). The SPAADIA Annotation Scheme. available from: http://martinweisser.org/publications/SPAADIA_Annotation_Scheme.pdf.

Searle, John. (1969). Speech Acts: an Essay in the Philosophy of Language. Cambridge: CUP.

Weisser, Martin. (2010). *Annotating Dialogue Corpora Semi-Automatically: a Corpus-Linguistic Approach to Pragmatics*. Habilitation (professorial) thesis, University of Bayreuth. 274 pages.

Weisser, Martin. (2014). Speech act annotation. In Aijmer, K. & Rühlemann, C. (Eds.). *Corpus Pragmatics: a Handbook*. Cambridge: CUP. Chapter DOI: 10.1017/CBO9781139057493.005.

Weisser, Martin. (2016). DART – the Dialogue Annotation and Research Tool. *Corpus Linguistics and Linguistic Theory.* DOI: 10.1515/cllt-2014-0051.

Weisser, Martin. (forthcoming 2016). Profiling Agents & Callers: a Dual Comparison Across Speaker Roles and British vs. American English. In Pickering, L., Friginal, E., & Staples, S. (Eds.). *Talking at Work: Corpus-based Explorations of Workplace Discourse*. London: Palgrave Macmillan.

Liste des auteurs