Book of Abstracts

-

Journée d'études AFLiCo JET

Corpora and Representativeness

jeudi 3 et vendredi 4 mai 2018
Bâtiment Max Weber (W)

Keynote speakers

Dawn Knight et Thomas Egan

Plus d'infos :
bit.ly/2FPGxub

Université Paris Nanterre

Université Paris Nanterre
–
EA 370
CREA
Centre de Recherches Anglophones

AFLiCo

modyco UMR 7114

UNIVERSITÉ PARIS 8 VINCENNES-SAINT-DENIS

Table of contents

DAY 1 – May 3rd 2018

Day 2 – May 4rd 2018

DAY 1 – May 3rd 2018

Thomas Egan (Inland Norway University of Applied Sciences)
thomas.egan@hihm.no

Some perils and pitfalls of non-representativeness

After some general introductory remarks on representativeness, I revisit in this paper some of the arena where I myself have encountered problems in connection with instances of non-representativeness, illustrating the discussion with examples from corpora both small and large, both historical and contemporary, and both mono-linguistic and contrastive. I discuss some text types that compilers of general corpora would do wise to steer clear of, before going on to describe some less immediately obvious phenomena that lie in wait to mislead the unwary researcher.

Daniel Henke (Université Paris 8 Vincennes Saint Denis)
dhenkel@univ-paris8.fr

De quoi sont représentatifs les corpus de textes traduits au juste ? Une étude de corpus comparable-parallèle

Dans la Stylistique Comparée de l'Anglais et du Français, Vinay et Darbelnet, cherchant à défendre et à définir la place de la science traductive parmi les sciences du langage, affirmaient « la traduction … pour observer le fonctionnement d'une langue par rapport à une autre, [est] un procédé d'investigation. Elle permet d'éclaircir certains phénomènes qui sans elle resteraient ignorés. À ce titre elle est une discipline auxiliaire de la linguistique (Vinay & Darbelnet, 1958 [Ns. soul.]). Depuis cette époque, la confrontation de textes-sources et cibles a pris une place centrale dans les études contrastives, notamment dans les courants s'inspirant des travaux de J. Guillemin-Flescher qui postule, dans sa préface au premier tome de la revue Linguistique Contrastive et Traduction, que « la comparaison de deux langues met à jour des phénomènes qui concernent le langage en général. (…) Elle constitue aussi un moyen de cerner de plus près les représentations qui caractérisent les deux langues envisagées. La mise en regard de textes traduits situe la réflexion linguistique dans une dimension contextuelle qui l'ancre d'emblée dans un enchaînement naturel. (…) l'examen attentif d'un corpus important met à jour des récurrences dans l'activité de traduction qui sont parfaitement systématisables » (Guillemin-Flescher, 1992 [Ns. soul.]). Le présupposé théorique et méthodologique qui sous-tend l'étude des textes traduits, aussi bien chez Vinay &Darbelnet que J. Guillemin-Flescher et leurs successeurs, pose toutefois un problème fondamentalde représentativité, car pour dire que l'étude de textes traduits permet d'« observer le fonctionnement d'une langue par rapport à une autre » et de « caractériser les deux langues envisagées », il faut d'abord admettre que les textes-cibles sont représentatifs de la langue-cible. Or, rien n'est moins certain. Depuis le milieu des années 80 s'est développé un autre courant depensée extrahexagonal qui voit dans la traduction tantôt un 'troisième code' (Frawley, 1984) distinctdes langues source et cible, tantôt une langue hybride (Gellerstam, 1986) aux traits métissés. Lerecours à des corpus 'comparables' constitués de textes originaux dans deux langues a été proposécomme remède, mais pose cette fois-ci le problème du degré de comparabilité (Laviosa, 1997).L'analyse proposée ici, laquelle s'appuie sur un corpus de 80 auteurs et traducteurs (20 auteurs anglophones, 20 traducteurs vers le français, 20 auteurs francophones et 20 traducteurs vers l'anglais) de 8 millions de mots environ, vise à montrer, à travers l'analyse statistique de plusieursindicateurs syntaxiques, que

les corpus de textes traduits manifestent bel et bien des traits hybrideset doivent ainsi être considérés représentatifs, non pas de la langue-cible, mais du processustraductif. Ce caractère hybride ne se voit généralement, toutefois, que dans la comparaison, non pasavec leurs textes sources, mais avec un autre corpus de textes non-traduits dans la langue-cible. Autrement dit, seule une approche croisée, menée à partir d'un corpus bilingue double, à la foiscomparable et parallèle, permet réellement « observer le fonctionnement d'une langue par rapport àune autre ».

Bibliographie

Frawley, W. (1984). "Prolegomenon to a theory of translation". In W. Frawley (ed.), in L. Venuti (ed.) 2000. *The Translation Studies Reader*, 250-263. London-New York: Routledge.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. Translation studies in *Scandinavia*, 1, 88-95.

Guillemin-Flescher, J. (Ed.). (1992). *Linguistique contrastive et traduction* (Vol. 1). Editions Ophrys.

Laviosa, S. (1997). How comparable can'comparable corpora'be?. *Target. International Journal of Translation Studies*, 9(2), 289-319.

Vinay, J. P., & Darbelnet, J. (1958). *Stylistique comparée de l'anglais et du français*. Paris/Montréal: Didier/Beauchemin.

Ilmari Ivaska, Silvia Bernardini, Adriano Ferraresi (Università di Bologna)
ilmari.ivaska@unibo.it; silvia.bernardini@unibo.it; adriano.ferraresi@unibo.it

The comparability paradox in multilingual and multi-varietal corpus research: Coping with the unavoidable

The present paper addresses the data comparability paradox – the dissonance between comparability and generalizability affecting the representativeness of corpora used in multilingual and multi-varietal corpus studies. We suggest that, by combining existing corpora with purpose-built ones, and exploring them through data-driven techniques, it might be possible to draw conclusions that are both reliable and generalizable across genres and language varieties, despite the complexity of the object of study.

While large-scale corpora are nowadays available in multiple languages, they often feature different genres and are, thus, not directly comparable. This is all the more true of varieties such as learner language (L2O), translated language (L1T), and non-translated first language (L1O). A study that aims to discover similarities and differences between these varieties would require a corpus with components as closely comparable as possible, in a set of genres and language pairs that guarantees generalizability: a near-impossible research object.

We have taken a dual approach to addressing this paradox. On the one hand, we assemble data opportunistically, making sure that they represent a multitude of (not-necessarily comparable) genres, and try to detect tendencies that set the studied varieties apart from each other. On the other, we collect a small but well-curated dataset in one genre for all the studied varieties, in all the studied languages, complementing it with survey data and interviews. The robust data-drivenness of the first phase can reveal general phenomena spanning across genres, varieties and languages, while the second phase helps refining findings and linking them to participants' perceptions about their own decision-making and about acceptability/naturalness of language use in general.

Our opportunistic corpus combines a variety of existing ones, and is complemented by ad-hoc texts collected from online sources. Our well-curated corpus consists of highly comparable opinion articles produced by early career language experts (advanced students or recent graduates in translation or language studies). We collect L1O, L1T, and L2O texts in three language pair combinations (English- Finnish, English-Italian, Italian-Finnish), as well as

acceptability judgements and qualitative interview data. The full dataset is parsed using the crosslinguistically consistent annotation scheme developed by the Universal Dependencies Initiative.

The analysis adopts an exploratory method – Key Structure Analysis – to reveal possible inter-varietal linguistic differences in a data-driven manner. N-gram frequencies (of lemmas, parts-of-speech, morphological forms, and syntactic functions) are treated as indicators of potential differences between the varieties, which are then further analyzed using random forests. Results are used as a point-of-departure to inform the research questions of the second phase.

Preliminary results suggest that it is possible to detect features that distinguish the studied varieties from each other, also disentangling genre-specific patterning from more general phenomena. We will study the detected constructions further in the second phase, with symmetrically comparable data. In sum, available resources can be used to lead the inquiry, as long as they are complemented with novel data to validate the generalizability of the results and to explore them in greater detail. This dual approach leads to interpretations that are robust in width without losing accuracy in depth.

Bibliography

Gries, Stefan Th. 2015. 'Some Current Quantitative Problems in Corpus Linguistics and a Sketch of Some Solutions'. *Language and Linguistics* 16 (1):93–117.

Ivaska, Ilmari. 2015. 'Longitudinal Changes in Academic Learner Finnish: A Key Structure Analysis'. *International Journal of Learner Corpus Research* 1 (2):210–41.

Ivaska, Ilmari, and Kirsti Siitonen. 2017. 'Learner Language Morphology as a Window to Crosslinguistic Influences: A Key Structure Analysis'. *Nordic Journal of Linguistics* 40 (2):225–53.

Lanstyák, Istvan, and Pál Heltai. 2012. 'Universals in Language Contact and Translation'. *Across Languages and Cultures* 13 (1):99–121.

Levshina, Natalia. 2017. 'A Multivariate Study of T/V Forms in European Languages Based on a Parallel Corpus of Film Subtitles'. *Research in Language* 15 (2):153–72.

Staples, Shelley, Jesse Egbert, Douglas Biber, and Susan Conrad. 2015. 'Register Variation. A Corpus Approach'. In *The Handbook of Discourse Analysis*, edited by Deborah Tannen, Heidi Ehernberger Hamilton, and Deborah Schiffrin, 2nd ed., 505–25. Oxford: Blackwell-Wiley.

Tagliamonte, Sali A., and R. Harald Baayen. 2012. 'Models, Forests and Trees of York English: Was/Were Variation as a Case Study for Statistical Practice'. *Language Variation and Change* 24:135– 78.

Antonina Bondarenko (Université Paris 7 Paris Diderot)
tonyabondarenko@gmail.com

Verbless Sentences: Advantages and Challenges of a Parallel Corpus-based Approach

In this paper, we discuss methodological issues facing corpus-based studies of verbless sentences. The difficulty of detecting the absence of the verb automatically has meant that most studies of verbless sentences have relied on fragmented data and syntactic theory has dominated the discussion (McShane 2000, Weiss 2013). Overcoming the typical problems of fixed annotation and verb-centric syntactic modeling associated with most existing parsed corpora (Landolfi et al. 2010), we develop an alternative method of automatic verbless sentence extraction using Trameur (Fleury & Zimina 2014) and evaluate its accuracy against manual data. Combining corpus methods with Guillemin-Flescher's (2003) principle that re-occurring translation patterns reveal linguistic constraints that would otherwise remain hidden, we investigate verbless sentences through parallel-text corpora in English and Russian with the aim of uncovering the semantic and pragmatic factors associated with the absence of the verb.

The present results are based on a pilot 150,000-word specially created corpus which includes Dostoevsky's dialogue-centered Russian *Brothers Karamazov* (1880), Pinter's English play *The Caretaker* (1960), and corresponding translations (Pevear & Volokhonsky 1990, Avsey 1994, Doroshevich 2006). Verbless sentences were automatically extracted, submitted to various statistical analysis, aligned with several translations, and analyzed in context. Furthermore, verbless utterances and their translation correspondences were manually annotated for antecedent-based ellipsis (McShane 2000), discourse type, information structure (Lambrecht 1994), and predication type (Hengeveld 1992). By creating a typology of segmentation and morphosyntactic errors, we were able to recall 95% of the verbless sentences automatically. Trameur permits automatic correction of a large portion of morphosyntactic tagging errors and makes it possible to visualize verbless sentences aligned with multiple translations in their original context.

Our results expose special technological requirements of the corpora used for verbless sentence studies, including: a) specific sentence segmentation that must be capable of distinguishing direct speech from narration, b) heightened accuracy for morphosyntactic tagging, and c) software capable of classifying the specially annotated and segmented sentences into those with the feature of a verb and those without. Parallel-text alignment poses additional challenges for sentence-level segmentation.

The results suggest that a particularly large corpus is necessary for both monolingual and parallel-text analysis of verbless sentences. While the pilot corpus allowed the computation of statistically characteristic elements of verbless sentences, the size proved limiting for repeated segments calculation. Translation patterns, revealed in bi-directional parallel-text analysis, were statistically limited both by the size of the present corpus and the necessity of manual annotation of all verbal sentences.

Furthermore, a statistically strong association of verbless sentences with direct speech (only 7% were found in literary narration) suggests that the corpus design of a verbless sentence study must consist of specially segmented corpora, and, in so far as the study aims for a high frequency of the phenomenon, it should privilege dialogue-centered texts and spoken corpora.

An ideal corpus for the contrastive study of verbless sentences should not only address compositional issues of corpus-based studies (Stolz 2007) and the pitfalls of parallel-corpora in general (Nádvorníková 2017), but must also take account of the technological biases facing verbless sentence corpus-based studies.

Bibliography

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

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

Hengeveld, K. 1992. Non-verbal Predication: Theory, typology, diachrony. Berlin: Mouton de Gruyter. Kopotev, M. (2007). Where Russian Syntactic Zeros Start: Approaching Finnish? In J. Nuorluoto (Ed.), *Slavica Helsingiensia*, 32, 116-137.

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

Landolfi, A., Carmela, S., & Voghera, M. (2010). Verbless clauses in Italian, Spanish and English: A Treebank annotation. In S. Bolasco et al. (Eds.), *JADT 2010: Proceedings of the 10th International Conference on Statistical Analysis of Textual Data* (pp. 1187-1194). Rome: CISU. (http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010- 1187-1194_066-Landolfi.pdf)

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

Nádvorníková, O. (2017). Pièges méthodologiques des corpus parallèles et comment les éviter. *Corela* [Online], HS-21, 2017. (http://corela.revues.org/4810)
Stassen, L. (2013). Zero Copula for Predicate Nominals. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures* Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (http://wals.info/chapter/120)
Stolz, T. (2007). Harry Potter meets Le petit prince – On the usefulness of parallel corpora in crosslinguistic investigations. *STUF-Sprachtypologie und Universalienforschung*, 60(2), 100-117.
Weiss, D. (2013). The lazy speaker and the fascination of emptiness: Colloquial Russian from a typological perspective. In Irina Kor Chahine (Eds.), *Current Studies in Slavic Linguistics*, 91-123. Amsterdam: John Benjamins Publishing Company.

Adeline Terry (Université Lyon III)
adeline.terry@univ-lyon3.fr

The representativeness of the metaphors of death, disease, and sex in a TV show corpus

The aim of this presentation is to study the metaphors for death, disease, and sex in a corpus constituted of the first two seasons of five American TV shows (*Six Feet Under, House, M.D, Grey's Anatomy, Sex and the City, How I Met Your Mother*). The main research question will be related to the collection and representativeness of data in such a spoken corpus.
Metaphorical expressions in the corpus were identified thanks to the Metaphor Identification Procedure (MIP), which was established by the Pragglejaz Group [2007]. The 423 metaphorical occurrences (137 metaphorical expressions for sex, 127 for death, and 159 for disease) were then classified according to their source domain using the framework of the Conceptual Metaphor Theory (first defined by Lakoff and Johnson [1980]). The methods used for the collection of the data will first be commented upon as both the videos and the scripts were used.
This data will then be analyzed and compared to the results found by several studies in which the metaphors of sex (Crespo Fernández [2008, 2017]), death (Crespo Fernández [2006], Kuczok [2009], Jamet [2010]), and disease (Jasen [2009], Semino et al. [2015, 2016, 2017]) were analyzed. The main objective will be to determine if the metaphors found in TV shows are similar to and representative of those found in natural conversation – notably regarding source domains. Quaglio [2009: 150] conducted a linguistic study in which he compared the language of the American TV show Friends to natural conversation, and concluded that they shared many characteristics and that TV shows therefore might provide significant data for linguistic analysis. However, he did not focus on metaphors but rather on conversational analysis. The hypothesis is that the source domains in our corpus will be quite similar in terms of both type and proportion to those found in natural conversation, even though there might be variations which will have to be accounted for.

Bibliography

BUCARIA, Chiara and Luca BARRA (ed.), *Taboo Comedy: Television and Controversial Humour*, Palgrave Studies in Comedy, 2016.
CRESPO FERNANDEZ, Eliecer, "Metaphor in the euphemistic manipulation of the taboo of sex", in *Babel-afial 15*, 2006(1) : 27-42.
CRESPO FERNANDEZ, Eliecer, "The Language of Death: Euphemism and Conceptual Metaphorization in Victorian Obituaries", in *SKY Journal of Linguistics* 19, 2006(3) : 101-130.
CRESPO FERNANDEZ, Eliecer, "Sex-Related Euphemism and Dysphemism: an Analysis in Terms of Conceptual Metaphor Theory", in *Atlantis, Journal of the Spanish Association of Anglo-American Studies* 30.2, 2008 : 95-110.

CRESPO FERNANDEZ, Eliecer, *Sex in Language: Euphemistic and Dysphemistic Metaphors in Internet Forums*, Bloomsbury, 2017.
DEIGNAN, Alice, *Metaphor and Corpus Linguistics*, John Benjamins Publishing Company, 2005.
DEMMEN, Jane, SEMINO, Elena, DEMJÉN, Zsófia, KOLLER, Veronika, HARDIE, Andrew, RAYSON, Paul et Sheila PAYNE, "A computer-assisted study of the use of violence metaphors for cancer and end of life by patients, family carers and health professionals", International *Journal of Corpus Linguistics* 20:2, John Benjamins Publishing Company, 2015: 205-231.
DEMJÉN, Zsófia, "Laughing at cancer: humour, empowerment, solidarity and coping online", *Journal of Pragmatics* 101, 2016: 18-30.
DEMJÉN, Zsófia, SEMINO, Elena, et Veronika KOLLER, "Metaphors for 'good' and 'bad' deaths: a health professional view", in *Metaphor and the Social World* 6:1 (2016), John Benjamins Publishing Company, 2016: 1-19.
JAMET, Denis, "Euphemisms for Death: Reinventing Reality through Words", dans SORLIN, Sandrine (ed.), *Inventive Linguistics*, Presses Universitaires du Languedoc et de la Méditerranée, Collection "Traverses", 2010.
JASEN, Patricia, « From the "silent killer" to the "whispering disease": ovarian cancer and the uses of metaphor », in *Medical history*, Vol. 53, No. 4, 2009: 489-512.
KÖVECSES, Zoltán, *Language, Mind and Culture*, Oxford University Press, 2006.
KUCZOK, Marcin, "Metaphorical conceptualizations of death and dying in American English and Polish: a corpus-based contrastive study", *Linguistica Silesiana* 37, 2016: 125-142.
LAKOFF, George and Mark JOHNSON, *Metaphors We Live By*, The University of Chicago Press, 1980.
PRAGGLEJAZ GROUP, "MIP: a method for identifying metaphorically used words in discourse", in *Metaphor and Symbol* 22(1), 2007 : 1-39.
QUAGLIO, Paulo, *Television Dialogue: the Sitcom Friends vs. Natural Conversation*, John Benjamins Publishing Company, 2009.
SEMINO, Elena and Zsofia DEMJEN (ed.), *The Routledge Handbook of Metaphor and Language*, London: Routledge, 2017.
SEMINO, Elena and Zsofia DEMJEN, *Metaphor, Cancer and the End of Life: A Corpus- Based Study*, Routledge: New York and London, 2018.

Julien Perrez, Pauline Heyvaert, Min Reuchamps (Université de Liège, Université Catholique de Louvain)
Julien.Perrez@ulg.ac.be; pheyvaert@uliege.be; Min.Reuchamps@uclouvain.be

On the representativeness of political corpora in linguistic research

There is a long tradition of linguistic research on political discourse from various theoretical perspectives, including critical discourse analysis (see among many others Fairclough 1995, Fairclough & Fairclough 2012, Wodak 1989), lexicometric approaches (see for instance Arnold 2005, Mayaffre 2005, 2016, Mayaffre & Poudat 2013, Authors 2015a) or cognitive linguistic approaches to metaphor (see among many other Charteris Black 2011, Musolff 2004, 2013, 2016 L'Hôte 2012). In these studies, political corpora collected from discourses by political elites (presidential debates, presidential addresses, public speeches,...) often appear to be overrepresented, leaving aside other forms of political discourses such as media discourse on political issues (see however Musolff 2004, 2013) or citizen discourse. As Bougher (2012 :149) posits for metaphor analysis: "while research on metaphors in political discourse has flourished in recent years, the focus on elite communication has left metaphor's wider capacity as a reasoning tool for citizens underexplored". This results in a certain lack of representativeness of the political domain in linguistic studies. Indeed, political discourse is not restricted to the political elites alone.

Advocating a more global to political corpora, including corpora from different subdomains of the political spectrum, our talk is structured in two main parts. Firstly, we will propose a quantitative bibliographic analysis aiming at assessing what type of political corpora are frequently used in linguistic research. Secondly, on the basis of previous and current analyses of different kinds of political corpora (including citizen, media and elite discourse) we have been collecting in the framework of the ADAPOF-project (see for example Authors 2015b), we will illustrate how taking this variety of political genres into account, allows us to unravel phenomena such as conceptual alignment or metaphor circulation, related to specific political issues (in this case Belgian federalism).

Bibliography

Arnold, E. (2005). Le discours de Tony Blair (1997-2004). *Corpus* 4 | URL : http://corpus.revues.org/340.

Authors (2015a). Constructing Quebec and Wallonia How political parties speak about their region. In Authors (ed), *Minority Nations in Multinational Federations : A comparative study of Quebec and Wallonia*, London & New-York : Rouledge, pp. 49-81.

Authors (2015). Folle machine ou solide relation « living apart together » ? Le rôle des métaphores dans la perception citoyenne du fédéralisme belge. *Mots. Les Langages du Politique*, 109 | URL : http://journals.openedition.org/mots/22156.

Bougher, L. (2012). The case for metaphor in political reasoning and cognition. *Political Psychology*, 33 (1), 145-163.

Charteris-Black, J. (2011). *Politicians and Rhetoric. The Persuasive Power of Metaphor*, Houndmills : Palgrave Macmillan.

Fairclough, N. (1995). *Critical Discourse Analysis : the Critical Study of Language*. New- York : Longman Group.

Fairclough, I. & Fairclough, N. (2012). *Political Discourse Analysis. A Method for Advanced Students*. London & New-York : Routledge.

L'Hôte, E. (2012). "Breaking up Britain"? Métaphores et discours sur la dévolution au Royaume-Uni. In Authors (eds.) *Les relations communautaires en Belgique, Approches politiques et linguistiques* , Louvain-la-Neuve, Academia- L'Harmattan (Science politique), p. 161-189.

Mayaffre, D. (2005). Les corpus politiques : objet, méthode et contenu. Introduction, *Corpus*, 4 | URL : http:// corpus.revues.org/292.

Mayaffre, D. (2016). Du candidat au président : Panorama logométrique de François Hollande. *Mots. Les Langages du Politique*, 112 | URL : http://journals.openedition.org/mots/22479.

Mayaffre, D. & Poudat, C. (2013).Quantitative Approaches to Political Discourse : Corpus Linguistics and Text Statistics. . In K. Fløttum (ed.), *Speaking of Europe : Approaches to Complexity in European Political Discourse*, Amsterdam & Philadelphia : John Benjamins, pp. 135-150.

Musolff, A. (2004). *Metaphor and Political Discourse Analogical Reasoning in Debates about Europe*. Houndmills : Palgrave MacMillan.

Musolff, A. (2013).The heart of Europe : Synchronic Variation and Historical Trajectories of a Political Metaphor. In K. Fløttum (ed.), *Speaking of Europe : Approaches to Complexity in European Political Discourse*, Amsterdam & Philadelphia : John Benjamins, pp. 135-150

Musolff, A. (2016). *Political Metaphor Analysis : Discourse and Scenarios*. London & New- York : Bloomsbury.

Wodak, R. (ed.)(1989). *Language, Power and Ideology : Studies in Political Discourse*. Amsterdam & Philadelphia : John Benjamins.

Joshua M. Griffiths (The University of Texas at Austin/Université Paris Nanterre)
grifjo06@gmail.com

Supplementing Maximum Entropy Phonology with Corpus Data

In their initial proposal for a Maximum Entropy model of phonology, Johnson & Goldwater propose that a learning algorithm of phonology should be able to "learn from a corpus of real, potentially noisy, data." (Johnson & Goldwater 2003:111.) Although much of the work in this theoretical framework has looked at proposing and testing the best learning algorithms and learning models, many of these analyses have relied on small corpora and even impressionistic data. In a more recent study, Bayles, Kaplan, and Kaplan (2016) have found that Harmonic Grammar is well-suited at describing phonologically variable phenomena and does adapt well to using data from large oral corpora. For this talk, I propose an analysis of obstruent-liquid schwa (OL) sequences in French such as those presented in (1). These structures have three possible phonetic outputs: either the liquid and schwa are deleted, only the schwa is deleted, or the liquid and schwa are preserved.

(1) a. une table grise [yn.ta.blə.gʁiz] ~ [yn.tab.gʁiz] ~[yn.tabl.gʁiz]
b. un arbre pourri [ɛ̃.naʁ.bʁə.pu.ʁi] ~ [ɛ̃.naʁb.pu.ʁi] ~ [ɛ̃.naʁbʁ.pu.ʁi]

In order to explain OL-schwa sequences in French I adapt a theory of weighted constraints (Pater 2009, Zuraw & Hayes 2017) in a maximum entropy (MaxEnt) grammar. A MaxEnt model is a probabilistically-driven machine learning model that weights constraints based on the probabilities of any possible output of the structure. Data come from the *Phonologie du Français Contemporain* corpus (Durand, Laks, Lyche, 2009), which is coded specifically for schwa realization and deletion. There are other factors that undoubtedly contribute to the realization of either possible output, which are explored through regression analysis. The regression analyses show that region is a significant predictor of schwa deletion; therefore, four grammars were constructed based on geographical region (Northern France, Southern France, Canada, Africa.) Despite using the same constraints, each variety of French had a different weighting condition of the constraint set. Southern French differed the most appearing to prefer phonological faithfulness to markedness. Future research will further incorporate the results of the regression analyses in conjunction with the MaxEnt grammars in a more comprehensive manner.

Key Words: Phonology, Variation, Schwa, Constraint-Based Grammar

Bibliography

Bayles, A., Kaplan, A., & Kaplan, A. (2016). Inter-and intra-speaker variation in French schwa. *Glossa: a journal of general linguistics*, 1(1).

Dell, F. (1976). Schwa précédé d'un groupe obstruante-liquide. *Recherches linguistiques Saint-Denis*, (4), 75-111.

Durand, J., Laks, B., & Lyche, C. (2009). Le projet PFC (phonologie du français contemporain): une source de données primaires structurées. *Phonologie, variation et accents du français*, 19-61.

Goldwater, S., & Johnson, M. (2003, April). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory* (Vol. 111120).

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive science*, 33(6), 999-1035.

Zuraw, K., & Hayes, B. (2017). Intersecting constraint families: an argument for Harmonic Grammar. *Language*, 93(3), 497-548.

Emmanuelle Guérin, Olivier Baude (Université d'Orléans, Université Paris Nanterre)
emm.guerin@gmail.com, olivier.baude@univ-orleans.fr

Représenter la variation – Revisiter les catégories et les variétés dans le corpus ESLO

La variation de la langue est souvent pensée selon la corrélation de « variétés » de langue (Gadet, 2012) et d'une caractérisation sociodémographique des locuteurs à laquelle se superpose la distinction entre ce qui est défini comme relevant de situations de communication formelles (qui donneraient lieu à un parler contrôlé) et informelles (qui donneraient lieu à un parler spontané).
Ainsi, la plupart des corpus, constitués dans une perspective variationniste, sont édifiés à partir d'une catégorisation sociale des locuteurs (âge, sexe, profession, niveau de scolarisation) et une catégorisation du « contexte », par le degré de formalisme de l'échange enregistré, dans la lignée des propositions de Labov (1972).
S'il n'est nullement question de remettre en cause ces catégorisations, nous cherchons à montrer qu'elles peuvent être insuffisantes pour une approche affinée de la variation de la langue (Guerin, 2017). Autrement dit, on envisage un second type de descripteurs, qui se place à un niveau moins objectif, pour cibler la particularité du contexte de production des données.
L'observation des données dans un grand corpus (en l'occurrence nous nous intéressons au corpus ESLO, (Baude & Bergounioux, 2016) met en évidence que les phénomènes retenus pour caractériser une « variété » en particulier n'apparaissent pas de façon systématique lorsque l'on a affaire aux catégories de locuteurs et de situations qui lui sont associées. On s'aperçoit que des facteurs relatifs à la spécificité de la situation d'interaction à un niveau micro, peuvent atténuer ou amplifier l'influence de facteurs d'un niveau macro. En fait, il s'agit de considérer « la réalité, telle qu'elle est (re)construite et interprétée par les sujets eux-mêmes » (Lüdi & Py, 1995 : 95)
Dans cette contribution, nous montrons comment cette approche théorique de la variation en deux niveaux imbriqués nous conduit à penser des orientations méthodologiques qui invitent à reconsidérer la représentativité des données du corpus. Dans cette perspective, on défend l'idée d'une réorganisation du corpus. Si on est en mesure de s'appuyer sur la caractérisation des situations aux deux niveaux imbriqués pour organiser la constitution d'un corpus pour représenter la variation, la validation des regroupements de données obtenues passe, in fine, par la confrontation aux productions. Cela implique qu'on envisage que l'organisation puisse évoluer au fur et à mesure du recueil. Le propre de la variation étant son inévitable dynamisme, Il revient au chercheur de ne pas chercher à la figée dans une modélisation qui, de fait, ne permettrait pas sa représentation.

Bibliographie

Baude, O., Bergounioux, G., (2016), chapitre « L'ESLO : une enquête en son temps » in Bergounioux, G., *Linguistique de corpus : une étude de cas La recette de l'omelette dans l'enquête socio-linguistique à Orléans (ESLO)*, Champion, Paris, 17-34.
Gadet F. (2012), « La variation, des variétés à la variabilité » in M. Dreyfus & J.-M. Prieur, *Hétérogénéités et variation. Perspectives sociolinguistiques, didactiques et anthropologiques*, Paris : Michel Houdiard Editeur, 27-38.
Guerin E. (2017), « Éléments pour une approche communicationnelle de la variation » in H. Tyne, M. Bilger, P. Cappeau et E. Guerin, *La variation en question(s)*, Bruxelles : Peter Lang, 57-76.
Labov W. (1972) (trad.1977), *Sociolinguistique*, Paris : Minuit.
Lüdi G.& Py B. (1995), *Changement de langage et langage du changement*, Lausanne : L'Age d'Homme.

Caroline Rossi, Camille Biros, Aurélien Talbot (Université Grenoble Alpes)
caroline.rossi@univ-grenoble-alpes.fr; camille.biros@univ-grenoble-alpes.fr; aurelien.talbot@univ-grenoble-alpes.fr

La variation terminologique en langue de spécialité : pour une analyse à plusieurs niveaux

Le plus souvent l'individu représente un niveau de granularité qui n'est pas conservé dans les grands corpus de référence, qu'il soit locuteur dans le cas de corpus oraux ou scripteur, auteur d'un texte intégré au corpus (voir par ex. Gries, 2015 : 99). Les linguistiques de corpus, pourtant issues d'une tradition contextualiste, semblent avoir délaissé les notions firthiennes de contexte de situation et de communauté de discours (Léon, 2008 : 28), en particulier avec l'avènement de bases de données de plus en plus volumineuses et l'établissement de conventions stables pour l'élaboration de corpus de référence. Quelle peut alors être la part des dimensions sociolinguistiques dans un corpus représentatif ? Cette question se pose avec d'autant plus d'acuité que nous assistons, depuis quelques années, au développement d'une linguistique variationniste basée sur corpus : Szmrecsanyi (2017 : 4) recense par exemple une vingtaine de publications dont plus de la moitié ont été publiées il y a moins de 5 ans.
Pour y répondre, nous nous intéressons aux apports d'une analyse à plusieurs niveaux des données de corpus (Gries, 2015 ; Schaefer, 2017), et en particulier à l'effet des facteurs nichés que sont les locuteurs et les types de textes. Les données sont des corpus écrits d'anglais de spécialité et la variation observée concerne des unités terminologiques complexes.
A la différence des corpus oraux, les corpus écrits sur lesquels nous travaillons se composent de textes le plus souvent produits par plusieurs auteurs. Nous commençons par définir les communautés d'auteurs qui structurent de telles données en termes de communauté de discours, ce qui nous permet de préciser notre question de recherche : les corpus spécialisés, qui représentent un genre ou un domaine (Kübler, 2011), permettent-ils la prise en compte de la variation liée à l'existence, en leur sein, de communautés de discours ?
Après avoir montré l'éclairage qu'apportent, pour l'analyse de la variation terminologique, les différents niveaux d'interrogation du corpus d'anglais médical Scientext dans l'outil ScienQuest (Falaise, Tutin & Kraif, 2011), nous nous concentrons sur un corpus « maison » de discours sur l'environnement (Auteurs, 2017). Les unités terminologiques analysées se rattachent au traitement des questions de justice, d'équité et de droits de l'homme dans ce corpus, qui contient des rapports d'organisations onusiennes, d'ONG et d'entreprises du secteur énergétique. Nous rapportons chacun des termes recensés à un type d'organisation (ONU, ONG, entreprise), et en son sein, à une communauté de discours, puis procédons à une modélisation statistique de la variation observée. Nous montrons à l'aide d'une régression multiple que les communautés de discours constituent le meilleur niveau de granularité.
Nos conclusions portent sur les implications de ces résultats, et nous insistons sur l'importance de constituer des bases de données aux unités « atomisables » (Habert, 2000) et ce jusqu'au plus fin niveau de granularité.

Bibliographie

FALAISE, A. TUTIN, A. & KRAIF, O. (2011). « Une interface pour l'exploitation de corpus arborés par des non informaticiens : la plate-forme ScienQuest du projet Scientext », Revue *TAL*, 52 :3, 103-128.
GRIES S. T. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* 10:1, 95-125.
HABERT, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment. In Bilger, M. (ed.) : *Linguistique sur corpus. Etudes et réflexions*, (31), 11–58, Perpignan : Presses Universitaires de Perpignan.
KÜBLER, N. (2011). Traduction pragmatique, linguistique de corpus, traducteur : un ménage à trois explosif ? *Tralogy*, session 3.
(http://lodel.irevues.inist.fr/tralogy/index.php?id=288&format=print)

LEON, J. (2008). Aux sources de la « Corpus Linguistics » : Firth et la London School. *Langages*, 171,(3), 12-33.
SCHAEFER, R. (2017) Mixed-effects regression modelling, in Gries & Paquot (eds.), *Practical Handbook of Corpus Linguistics.*

DAY 2 – May 4th 2018

Dawn Knight (School of English, Communication and Philosophy, Cardiff University)
KnightD5@cardiff.ac.uk

Representativeness in CorCenCC: corpus design in minoritised languages

Corpus design and construction in a minoritised language context pose interesting challenges, but also present opportunities not always open/available to developers of corpora for larger languages. During this presentation, I will discuss and examine these challenges and opportunities in more detail, with reference to the design and construction of the ESRC/AHRC-funded CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – National Corpus of Contemporary Welsh) corpus. Its ambition to create a large-scale open-source corpus of contemporary Welsh, one with a functional design informed, from the outset, by representatives of all engaged academic and community user groups, makes CorCenCC a highly relevant point of reference for lines of discussion that will be presented.
The construction of CorCenCC began in 2016 and, when complete in 2019, it will be the first *general* corpus of Welsh language. It will include data from a range of different discourse contexts (from formal contexts, e.g. political documents, televised interviews and formal letters, to less formal ones, e.g. informal emails, phone calls and text messages), and geographical locations in Wales. Data will also be sampled from a range of different speakers and users of Welsh, so from all regions of Wales, of all ages and genders, with a wide range of occupations, and with a variety of linguistic backgrounds (e.g. how they came to speak Welsh), to reflect Wales' diversity not only of text types but also of Welsh speakers themselves. This composition will allow users to make generalised observations about language use (i.e. not restricted to a specific discourse context or domain). CorCenCC will contain 10 million words by the end of the project, comprising 4 million each from spoken and written discourse and 2 million from digitally mediated discourse (e-language).
Defining and maintaining 'balance' and 'representativeness' in the design and construction of CorCenCC, and ensuring that all contemporary speakers and users in language communities that are located in areas and domains with different densities of both users and usage, is a major challenge. This is something compounded further when utilising unplanned, spontaneous contributions such as those derived from crowdsourcing means, which is something that the data collection process for CorCenCC is pioneering. These challenges will be unpacked within this presentation. Particular attention will be given to how we are ensuring that the design frame for CorCenCC truly reflects current social, cultural, geographical elements bearing upon contemporary Welsh as well as the domains in which it is used: providing guidelines for good practice that can be adapted to other minoritised language contexts. Metadata plays a key role in organising the ways in which a language corpus can be meaningfully analysed, so I will also demonstrate a streamlined, searchable database system for recording information about data collection and metadata as part of this discussion.
The presentation will end with an examination of some of the potential applications of CorCenCC and how priorities and engagement within a minoritised language context differ from those in major languages.

Frederick Newmeyer (University of Washington)
fjn@uw.edu

Conversational corpora: When 'big is beautiful'

The goal of this presentation is to examine the relationship between corpus size and conclusions drawn from corpora regarding questions of grammatical theory. Many linguists, typically those with a 'usage-based' orientation, question whether introspective data, as opposed to corpus-derived data, is or can be relevant to the construction of the correct theory of language. In their view, the 'disembodied sentences that analysts have made up ad hoc, ... rather than utterances produced by real people in real discourse situations' (Tomasello 1998: xiii) lead inevitably to the supposedly unrealistic complex abstract structures posited by generative grammarians. Usage-based grammarians assert that if one focuses on naturally occurring discourse drawn from corpora, then grammar will reveal itself to be primarily a matter of memorized formulas and simple constructions. This paper challenges that view. Appealing to a 170MB corpus of conversational English, it argues that introspective data and corpus- derived data do not lead to different conclusions about the nature of linguistic theory.

I discuss two publications that appear to back up Tomasello's claim. The first is Thompson & Hopper (2001), which attempts to show that the notion of 'argument structure', which is central to most formal theories, is irrelevant to language. They argue that '[T]he apparent importance of [argument structure] may be an artifact of working with idealized data. Discussions about argument structure have to date been based on fabricated examples rather than on corpora of ordinary everyday talk' (Thompson & Hopper 2001: 40). The second is the book *Spontaneous Spoken Language* (Miller & Weinert 1998). Miller & Weinert conclude that '[t]he properties and constraints established over the past thirty years by Chomskyans [are based on sentences that] occur neither in speech nor in writing [or only] occur in writing' (Miller & Weinert 1998: 379). Among the examples of English sentences that are supposedly never spoken in conversation, Thompson & Hopper and Miller & Weinert cite backwards anaphors ('Because she is smart, Mary will succeed'), gapping ('I ordered the chicken and Mary the fish'), and gerunds with auxiliaries ('Because of my having lived in France, I can speak the language').

The problem is the tiny size of the corpora that Thompson & Hopper and Miller & Weinert use. The former paper is based on only 446 clauses from three face-to-face multi-party conversations and the latter book limits itself to an English corpus of only 50,000 words. In my 170MB corpus, the Fisher English Training Transcripts, all of the above constructions are found and many more that have been claimed to occur only in writing. In other words, corpus size makes a big difference.

My conclusion is not to reject the use of conversational corpora, which are extremely valuable. Rather, it is to argue that if one wishes to draw conclusions about grammatical theory from corpora, then 'big is (indeed) beautiful'. When large corpora are used, then introspective data and corpus-derived data do not lead to different conclusions about the nature of linguistic theory.

Bibliography

MILLER, JIM, AND REGINA WEINERT. 1998. *Spontaneous spoken language: Syntax and discourse.* Oxford: Clarendon.

THOMPSON, SANDRA A., AND PAUL J. HOPPER. 2001. *Transitivity, clause structure, and argument structure: Evidence from conversation. Frequency and the emergence of linguistic structure*, ed. by J. L. Bybee, and P. Hopper, 27-60. Amsterdam: John Benjamins.

TOMASELLO, MICHAEL (ed.) 1998. *The new psychology of language: Cognitive and functional approaches to language structure*. Mahwah, NJ: Lawrence Erlbaum.

Graham Ranger (Université d'Avignon et des Pays de Vaucluse)
Graham.Ranger@univ-avignon.fr

How to get "along": in defence of an enunciative and corpus-based approach

The representativeness of a corpus is a function of the relationship between the corpus and a target language or language variety the corpus is intended to sample. It is however impossible to assess this relationship satisfactorily since its second term -- the target language -- can only ever be apprehended via a finite set of language occurrences, i.e. a corpus. Given this difficulty it is perhaps simpler and more relevant to consider representativeness with respect to the specific research goals one sets oneself.
In the enunciative approach to language, markers are described in terms of an invariant schematic form which is configured by the pressure of surrounding operations into contextually-situated shapes. The schematic form is formulated in terms of a limited number of operands and operations. The elaboration of a schematic form is carried out on the evidence of authentic examples studied in context and submitted to manipulations and judgements of acceptability on the basis essentially of the intuition of the linguist. In this presentation I will argue that corpus methodologies can provide invaluable quantitative support for these intuitions and, from there, for the theoretical constructions of enunciative approaches to language. I aim to argue this point with a case study of the marker "along", using the data of the British National Corpus.
More specifically, corpus queries will allow us to isolate a number of typical occurrences of "along". From this starting point, I will examine the role played by surrounding context in parametering the operation marked by "along". The collocational affinities of "along" will allow us progressively to identify four particularly significant contextual configurations. These yield spatial, temporal, subjective and argumentative values for the marker. In each type of occurrence, it is claimed that "along" marks an operation of identification between a locatum and a locator, defined as an unbounded, sequentially ordered space. This model will also be shown to help in characterising the compound form "alongside", the cluster "along with" or the suffixal use of "-along".

Bibliography

Culioli, Antoine. 1990. *Pour une linguistique de l'énonciation: opérations et représentations*. (Collection L'Homme Dans La Langue). Gap, France: Ophrys.
Hoffmann, Sebastian, Stefan Evert, Nicholas Smith & David Lee. 2008. *Corpus linguistics with BNCweb: a practical guide. (English Corpus Linguistics v. 6)*. Frankfurt am Main: Peter Lang.
Lakoff, George. 2012. Wom*en, fire, and dangerous things: what categories reveal about the mind*. paperback ed., [Nachdr.]. Chicago: The Univ. of Chicago Press.
Langacker, Ronald W. 2000. *Grammar and conceptualization. (Cognitive Linguistics Research 14)*. Berlin ; New York: Mouton de Gruyter.
Ranger, Graham. 2014. " Fire away " Suggestions pour une caractérisation énonciative de la particule adverbiale anglaise AWAY. In Jean-Marie Merle (ed.), *Faits de langues. Prépositions et aspectualité*. https://hal-univ- avignon.archives-ouvertes.fr/hal-01340029.
Tyler, Andrea & Vyvyan Evans. 2007. *The semantics of English prepositions: spatial scenes, embodied meaning and cognition.* Digitally printed version (with corr.). Cambridge [u.a]: Cambridge Univ. Press.

Thi Thu Trang Do, Huy Linh Dao (Université d'Orléans, INALCO)
dothithutrang12@gmail.com; dao.huy.linh@gmail.com

Corpus et représentativité : le cas de la concession en français parlé

Jusqu'à ce jour, la plupart des recherches consacrées à la concession ont opéré à partir d'exemples fabriqués ou de corpus écrits (cf. Moeschler & Spengler 1981, 1982 ; Létoublon 1983; Martin 1987; Le Pesant 2005, 2006; Soutet 2008, entre autres) en sorte que les phénomènes présents à l'oral sont ignorés ou minorés. Cette étude interroge les moyens d'expression de la concession en français parlé afin d'en appréhender les spécificités et le fonctionnement.

Pour mener à bien ce travail, nous avons commencé par faire des requêtes dans deux grands corpus oraux, ESLO et CLAPI. Parce que le premier est conçu dans une perspective sociolinguistique à partir d'enquêtes et le second contient principalement des interactions professionnelles, peu d'occurrences pertinentes pour une analyse pouvaient en être extraites. Il fallait envisager un corpus dédié. Quel genre de discours choisir ? Le discours des médias permettait de s'affranchir des questions juridiques et les débats apparaissaient comme une ressource potentielle du fait de l'interactivité et des échanges argumentatifs. Le choix d'émissions et de débats était essentiel. Huit émissions du *Grand Bûcher* de France Bleu Orléans, soit cent soixante et onze minutes de parole, ont été choisies. Il ne s'agit pas d'un corpus volumineux mais représentatif pour plusieurs raisons.

Diffusé cinq jours sur sept du lundi au vendredi, le *Grand Bûcher* est l'émission la plus décapante du paysage radiophonique orléanais. Cela a permis un large choix des émissions les plus adaptées à notre but, c'est-à-dire celles qui présentent beaucoup de concessions. Les présentateurs et les polémistes n'étant pas les mêmes d'une fois sur l'autre, nous pouvions donc disposer d'une grande variété dans les modalités d'expression et d'argumentation. De plus, les invités, qui peuvent être des experts du sujet à débattre, s'expriment de manière plus spontanée, ce qui nous offre quantité d'expressions argumentatives dont concessives. Cette ressource est d'autant plus exploitable que le langage utilisé dans ces débats n'est ni appauvri ni formaté, contrairement à ce que l'on observe dans les débats politiques où les locuteurs utilisent des formules stéréotypées. Comme nos corpus sont des corpus oraux situés transcrits par nous- mêmes, nous pouvons contrôler les données, les exploiter tout en ayant conscience du schéma intonatif qui les structure et du contexte situationnel qui détient tout le réseau de références sans lequel l'énoncé est privé de sa substance sémantique spécifique (cf. Morel 1996).

L'examen approfondi de nos corpus nous a permis de dégager un certain nombre de tendances : (i) *quoique*, d'ordinaire considéré comme le subordonnant prototypique de la concession à l'oral, est absent de notre corpus ; (ii) un tiers des concessions relevées porteraient essentiellement sur la valeur des mots, ce qui est présenté comme une renégociation de leur définition (catégorie que nous appelons *concession définitionnelle*) ; (iii) quand la formulation de ce type de concessions est exhaustive, elle comporte les cinq éléments suivants : le présentatif (PRES), la modalisation (MOD), le terme générique ou abstrait (TERME), la négation (NEG) et le connecteur à valeur concessive (CC). A l'oral, ces composants ne sont pas toujours des formes typiques mais des équivalents en discours.

Après avoir distingué entre ce type de concessions et celles qui discutent plutôt de la réalité des faits eux-mêmes, on montrera que l'affinement de l'analyse requis par l'expression de la concession tient pour une part à la diversité de ses formulations et plus encore à leur dissémination. L'absence des conjonctions (ou des prépositions) attendues se trouve compensée par la distribution, au fil du discours, d'une succession d'unités lexicales hétérogènes qui, par leur réunion, cadrent la séquence recherchée et en signalent la valeur. La récurrence des formes et leur repérage discontinu dans la linéarité de la parole proférée

permettent de formaliser une séquence- type maximale qui mobilise des éléments identiques, tantôt intégralement, tantôt partiellement.

Bibliographie

BAUDE, Olivier (2012). Corpus de référence : homogénéité, hétérogénéité et représentativité. *Journées ILF : Initiative Corpus de référence*.

KANG, Shin-Tae (2015). Représentativité d'un corpus de terrain : le choix du périmètre d'investigation. *ICODOC 2015 : Colloque Jeunes Chercheurs du Laboratoire ICAR*. [En ligne], SHS Web of Conference 20, consulté le 01 décembre 2017.
https://www.shs-conferences.org/articles/shsconf/pdf/2015/07/shsconf_icodoc2015_01012.pdf

LE PESANT, Denis (2005). Causalité et concession. *Questions de classification en linguistique: méthodes et description. Mélanges offerts à Christian Molinier*, 195-210.

LE PESANT, Denis (2006). De la concession à la cause, et de la cause à la condition. *La cause: approche pluridisciplinaire, Linx, 54*: 61-71.

LÉTOUBLON, Françoise (1983). Pourtant, cependant, quoique, bien que: dérivation des expressions de l'opposition et de la concession in Connecteurs pragmatiques et structures du discours. Actes du Colloque de pragmatique de Genève (7-9mars 1983). *Cahiers de Linguistique Française, 5*: 85-110.

MARTIN, Robert (1987). *Langage et croyance. Les « univers de croyance » dans la théorie sémantique*. Liège, Mardaga.

MOESCHLER, Jacques & SPENGLER, Nina (de) (1981). Quand même : de la concession à la réfutation. *Cahiers de linguistique française*, *2*: 93-112.

MOESCHLER, Jacques & SPENGLER, Nina (de) (1982). La concession ou la réfutation interdite. Approches argumentative et conversationnelle in Concession et consécution dans le discours. *Cahiers de Linguistique Française, 4*: 7-36.

MOREL, Mary-Annick (1996). *La concession en français*. Paris : Ophrys.

SOUTET, Olivier (2008). Des concessives extensionnelles aux concessives simples. Contribution à l'étude de la genèse sémantique et historique des locutions conjonctives concessives du français. *Linx, 59*: 115- 132.

Dominique Boutet, Claudia Bianchini, Claire Danet, Patrick Doan, Morgane Rébulard, Adrien Contesse, Léa Chèvrefils-Desbiolles (Université de Rouen, Université de Poitiers, ESAD Amiens, UTC Compiègne)
dominique_jean.boutet@orange.fr; claudia.savina.bianchini@univ-poitiers.fr; claire.danet@gmail.com; pdoan.atelier@gmail.com; morgane.rebulard@gmail.com; adriencontesse@gmail.com; leachevrefils@gmail.com

Handling Sign Language annotations of the handshapes

Historically, transcriptions in SL have largely focused on relaying the meaning of a sign rather than the sign's physical shape [1,2]. Two main graphical systems of annotation —SignWriting [3] and HamNoSys [4]— are able to annotate the form of the parameters: the first one is not designed to annotate [5]; the second, in most cases, is coupled to a database tool for integrating SL corpus, named iLex [6]. Therefore, except for few corpora [7, 8], the parameters are only partially annotated. One of the reasons is the time needed to do so, even if this has not been documented in literature. Another challenge in the transcription of signs is the integration of generic information (features) into the specificities of each parameter of signs.

At a phonological level [9, 10], the features extracted for handshapes generate multi-million possibilities. Some rules to reduce these potential combinations have been established [11, 12], and others might still be found.

The aim of this communication is to present Typannot [13, 14], a type font for handshapes created to transcribe all the existing SL (142 SL for Ethnologue [15]). This glyphic type font is readable, writable, searchable. It works upon a modular system; the separate features can be assembled into glyphs that

enable the representation of every handshape. The features are all searchable through a set of generic (symbolic) glyphs. Those glyphs can be visualized by the annotator at any time thanks to the use of ligatures (Opentype format [16]), and are therefore transparent. At present, this font —freely downloadable on every operating system— can be used with ELAN [17] to transcribe directly by using a virtual keyboard (MacOS) or, for some SL, by using a dedicated template.
Apart from the keyboard and templates for ELAN, a third annotation device has been developed. It works with a motion capture (MoCap) device (based on Leap Motion [18]) by which the annotator can transcribe a corpus by directly using his own hand (real-time controlled). We are in the process of measuring the annotation time per minute of corpus for the different levels and devices: with generic glyphs, with specific glyphs and with MoCap. The tendency seems to favour of the use of the MoCap device.
Finally, we will present some preliminary results of the transcription made with the Typannot Typefont, at several levels of annotation: from the selection of the fingers, the shapes and the bending angles of the fingers in the palm, their lateral distance to the final glyph of each handshape through the combination shape+angle.

Bibliography

[1] Johnston, T. (2008). Corpus linguistics and signed languages: no lemmata, no corpus. *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages* (Marrakesh, Morocco):82-88.
[2] Fenlon, J., Schembri, A., Johnston, T., & Cormier, K. (2015). *Documentary and corpus approaches to sign language research. Research Methods in Sign Language Studies: a Practical Guide*. Wiley-Blackwell, Hoboken (NJ, USA):156-172.
[3] Sutton, V. (1995). *Lessons in SignWriting: Textbook*. DAC, LaJolla (CA, USA).
[4] Prillwitz, S., Leven, R., Zienert, H., Hanke, T., & Henning, J. (1989). *Hamburg Notation System for sign languages: an introductory guide.* Signum Press, Hamburg.
[5] Bianchini, C.S. (2012). *Analyse métalinguistique de l'émergence d'un système d'écriture des Langues des Signes : SignWriting et son application à la Langue des Signes Italienne (LIS).* PhD dissertation (Univ. Paris8 & Univ. Studi Perugia).
[6] Hanke, T. (2002). iLex - A tool for sign language lexicography and corpus analysis. *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (Las Palma de Gran Canaria, Spain):923-926.
[7] Efthimiou, E., Fontinea, S.E., Vogler, C., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos P., & Segouat, J.(2010). Dicta-sign–sign language recognition, generation and modelling: a research effort with applications in deaf communication. *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* (La Valletta, Malta):80-83.
[8] Hanke, T., König, S., Konrad, R., & Langer, G. (2012). Towards tagging of multi-sign lexemes and other multi-unit structures. *Proceedings of 8th International Conference on Language Resources and Evaluation* (Istanbul, Turkey).
[9] Liddell, S. K. (1990). Structures for representing handshape and local movement at the phonemic level. *Theoretical Issues in Sign Language Research*, 1:37-65.
[10] Brentari, D. (1998). *A prosodic model of sign language phonology*. The MIT Press, Cambridge.
[11] Ann, J. (1996). On the relation between ease of articulation and frequency of occurrence of handshapes in two sign languages. *Lingua*, 98(1):19–41.
[12] Brentari, D. (2012). Phonology. In Pfau R., Steinbach M. & Woll B. (eds.), *Sign language: an international handbook.* Walter de Gruyter, Berlin: 21-55.
[13] Boutet, D., Bianchini, C. S., Doan, P., Goguely, T., Rébulard, M., & Danet, C. (2016, juillet). Typannot: a glyphic system for the transcription of handshapes. Colloque international présenté à 7th Conference of the International Society of Gestures Studies, Paris.

[14] Boutet, D., Doan, P., Danet, C., Bianchini C. S., Goguely, T., Contesse, A., Rébulard, M., (Forhtcoming). Systèmes graphématiques et écritures des langues signées. *Signata*, 8. Presses universitaires de Liège.
[15] Lewis, M.P. (2017). Ethnologue: languages of the world, vers. 20. SIL International. http://www.ethnologue.com [online; accessed: 1/12/2017]
[16] Wikipedia (2017). OpenType. https://en.wikipedia.org/wiki/OpenType [online; accessed: 1/12/2017].
[17] Crasborn, O., & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. Proceedings of 6th International Conference on Language Resources and Evaluation (Marrakesh, Morocco).
[18] Wikipedia (2017). Leap Motion. https://en.wikipedia.org/wiki/Leap_Motion [online; accessed: 1/12/2017].

Christophe Parisse (INSERM)
cparisse@u-paris10.fr;

How much coverage might a dense corpus provide?

Dense corpora have been put forward as necessary tools for corpus studies of language acquisition. Despite the great interest of this approach, it difficult to judge how large such a corpus should be. The idea put forward by Tomasello and Stahl (2004) is that a dense corpus should be large enough so that rare phenomena appear frequently enough to be studied correctly. Another purpose (cf. Lieven et al., 2003) is to explain how new material produced by a speaker can be explained on the basis of the previous language produced or heard.
In order to evaluate the quality of a corpus, I used the same principle and tried to measure how much a new utterance is covered by the corpus. If the utterance is covered at 100%, then the corpus is perfect. If the coverage is 50%, then the corpus is probably too small. The idea that I will test is whether large-size dense corpora reach 100%, and whether a part of the same corpora reach the same results.

Two measures will be used:
- **lexicon**: how many words in an utterance are in the corpus;
- **syntax**: how many word bigrams in an utterance are in the corpus.

Word bigrams are used to account for syntactic knowledge as in a construction grammar approach they form the basic information for categorization and generalization. So they should be informative enough to allow syntactic knowledge induction.

I used three corpus from the CHILDES database, the Thomas corpus (Lieven et al., 2009): 379 sessions, the Sarah corpus (Brown corpus: Brown, 1973): 139 sessions, and the Lily corpus (Providence corpus: Song et al., 2009): 80 sessions. For each session of each of the three corpora, the percentage of words (or bigrams) which are present in the recording and were found in the previous recordings will be computed, according to: 1) the number of previous recordings taken into account; 2) inclusion or not of the current session in the available knowledge.
Results showed that for lexical knowledge, coverage went up to 97% for the Thomas corpus after 40 sessions and did not increase with more sessions. Without including the current session, the maximum reached was 95%, and it was reached later after 70 sessions. Similar results were found for the other corpora, which means that the Sarah and the Lily corpus are in fact dense, although recordings are not as frequent as for the Thomas corpus.
Results for syntax were not as strong. Coverage never really stopped growing. However, the coverage seemed to stabilize at 75%, and was already at 70% after 50 sessions (more quickly for Lily, 40

sessions, more slowly for Sarah, 80 sessions). As much syntactic knowledge is never seen (at least 25 percent of the word bigrams), it is not clear how much interesting information would be provided by using corpus with more than 50-60 sessions.
So our results showed that corpora of at least 30 or 40 one-hour transcriptions were useful to study dense corpora, but longer corpora did not provide much more interesting information.

Bibliography

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: a usage-based approach. *Journal Of Child Language*, 30(2), 333–370.
Lieven, E., Salomo, D. & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20, 3, 481-508.
Song, Jae Yung, Sundara, Megha, & Demuth, Katherine. 2009. Phonological Constraints on Children's Production of English Third Person Singular -s. *Journal of Speech, Language, and Hearing Research*, 52(3): 623-642.
Tomasello, M. & Stahl, D. (2004). Sampling children's spontaneous speech: how much is enough? *Journal of Child Language*, 31(1), 101–121.

Guillaume Desagulier, Frédéric Isel, Anne Lacheret-Dujour, Seongmin Mun (Université Paris Nanterre, Ajou University)
gdesagulier@univ-paris8.fr; fisel@u-paris10.fr; anne.dujour.27@gmail.com; skdkflak@naver.com

Characterizing discourse genres with prosodic features in a reference treebank of spoken French

Rhapsodie is a 33000-word treebank of spoken French that is annotated for syntax and prosody. It breaks down into 57 five-minute long samples produced by 89 male and female speakers. The discourse profile of each sample is captured by six variables: event structure (dialogue vs. monologue), social context (public vs. private), genre (argumentation, description, narrative, oratory, and procedural), interactivity (interactive, non-interactive, and semi-interactive), channel (broadcasting and face-to-face), and planning type (planned, semi-spontaneous, and spontaneous).

The prosodic profile of each sample is captured by two sets of three variables. The first set consists of primary (i.e. structurally objective) variables, namely the mean number per second of pauses (fPauses), conversational overlaps (fOverlap), and gap fillers (fEuh). The second set is based on a model consisting of secondary variables determined a priori by the authors because they are likely to occur in certain discourse genres. They are the mean numbers per second of prosodic prominences (fProm), intonational periods (fIPE), intonation packages (fIPA).

Our main research question is whether discourse types in French can be characterized and ultimately predicted by prosodic features. We also address two side questions. First, does the fact that the corpus is relatively small, heterogeneous, and not necessarily balanced affect the representativeness of our results? Second, are the secondary prosodic features representative of discourse genres?
We compiled a data table that consists of 57 observations (the corpus samples) and the twelve above listed variables. We visualized the table with RhapVis, a tool we designed on purpose (Fig. 1), explored it with principal component analysis (Fig. 2), and looked for confirmed tendencies with non-parametric one-way ANOVAs (Kruskal-Wallis H tests).

Our exploration shows that argumentative and narrative sequences are prosodically marked, whereas descriptive and procedural sequences are not. A discourse genre is prosodically marked when it is characterized by a high frequency of prosodic features, namely the simultaneous occurrence of overlaps, prominences, and intonation packages. We also claim that a discourse genre

is prosodically marked when it is atypical with respect to the other speech genres. This is the case with oratory speech, which is characterized by a high frequency of intonational periods and pauses and is consequently isolated from the other types.

These results were partially confirmed by the ANOVAs. Focusing on primary variables, running an ANOVA on fPause showed a significant main effect of Genre ($p < 0.05$). Further inspection indicates that while the lowest fPause score was found in Narration ($M = 0.32$; $SD = 0.04$), the highest score was observed in Oratory ($M = 0.42$; $SD = 0.01$). For fOverlap, the main effect of Genre reached the level of significance ($p < 0.001$), indicating that fOverlap also varies according to Genre. The descriptive data showed that the fOverlap score was the highest for both Argumentation ($M = 0.05$, $SD = 0.04$) and Narration ($M = 0.02$, $SD = 0.01$). Conversely, no overlap was found in both Oratory and Procedural samples.

Bibliography

Lindqvist, Christina. *Corpus transcrits de quelques journaux télévisés français*, Stockholm, Elanders Gotab, 2001, 289 pages

Portele T, Heuft B, Widera C, Wagner P, Wolters M (2000) Perceptual Prominence In: *Speech and Signals. Aspects of Speech Synthesis and Automatic Speech Recognition. Festschrift dedicated to Wolfgang Hess on his 60th birthday. Forum Phoneticum*, 69. Hektor, Frankfurt a.M.: 97-116.

Wagner, P. et al. (2015b), « Disentangling and connecting different perspectives on prosodic prominence », Communication à ICPL, International Conference Prominence in Language, 2015, Cologne, ICPH, 2015

Hugo Chatellier et Cécile Viollain (Université Paris Nanterre)
hchatell@parisnanterre.fr; cviollain@parisnanterre.fr

Phonologie et petits corpus : un mariage de raison ?

Nous prenons pour point de départ de notre réflexion un constat simple : à l'heure où l'essor de la linguistique de corpus et le développement du *big data* ont multiplié les ressources, la survie et le recours toujours possible aux petits corpus ont surpris les experts intéressés par les enjeux épistémologiques et méthodologiques de cette discipline (McEnery & Wilson 2001 : 191). Cette survie est, selon nous, la preuve d'une nécessité à la fois matérielle et logistique et d'une pertinence scientifique des petits corpus que nous explorons sous l'angle particulier de la langue orale et du domaine spécifique de la phonologie. Dans un premier temps, nous revenons sur le rapport unique à l'objet langue qu'entretient la phonologie, par rapport à la syntaxe et à la morphologie notamment, étant donné sa non-récursivité (Scheer 2004). Nous définissons alors ce que nous entendons par « corpus phonologique », à savoir un outil privilégié pour l'étude quantitative et qualitative de phénomènes précis de la langue orale. Nous repensons sa validité non pas en termes de taille, mais en termes de finalité, à savoir le fait pour un corpus de se voir assigner un but précis par celui qui le construit, d'être pensé comme moyen pour atteindre une fin spécifique et donc de constituer un outil adapté à ce qu'il est censé chercher, et trouver. Nous essayons, par là-même, de dépasser l'inévitable aporie de la question de la représentativité des corpus en démontrant la pertinence individuelle de petits corpus phonologiques et en soulignant l'apport concret de leur comparabilité et de la jonction possible des analyses qu'ils permettent de formuler. Pour illustrer nos propos, nous utilisons les données des corpus PFC (Phonologie du Français Contemporain, Durand, Laks & Lyche 2009, Detey *et al*. 2016) et PAC (Phonologie de l'Anglais Contemporain, Brulard, Carr & Durand 2015). Plus précisément, nous synthétisons les résultats obtenus par ces deux programmes en ce qui concerne le(s) phénomène(s) de liaison, à savoir la liaison en français et le « r » de *sandhi* en anglais (Soum-Favaro, Coquillon & Chevrot 2014). Nous montrons comment ces deux programmes traitent de ces phénomènes à l'aide de codages qui permettent de rester le plus neutre possible d'un point

de vue théorique, et comment ils permettent de formuler des analyses quantitatives et qualitatives sur ces phénomènes à l'échelle de corpus individuels (corpus PAC Manchester et Nouvelle-Zélande pour l'anglais, et corpus PFC français laurentien (Canada) et français métropolitain (France) pour le français) mais également dans une perspective comparative (étude de la phénoménologie de la liaison propre à chaque langue). Aussi, en défendant les petits corpus, nous défendons en creux la phonologie de corpus, et nous remettons en cause la supériorité automatique des grands corpus pour ce qui est de l'analyse en profondeur de la langue orale.

Bibliographie

Brulard, I., Carr, P. & Durand, J. (dir.) (2015) *La prononciation de l'anglais contemporain dans le monde : Variation et structure.* Toulouse : Presses Universitaires du Midi.
Detey, S., Durand, J., Laks, B. & Lyche, C. (dir.) (2016) *Varieties of Spoken French*. Oxford : Oxford University Press.
Durand, J., Laks, B. & Lyche, C. (dir.) (2009) *Phonologie, variation et accents du français*. Paris : Hermès.
McEnery, T. & Wilson, A. (2001) *Corpus Linguistics: An Introduction*. 2ème édition. Edimbourg : Edinburgh University Press.
Scheer, T. (2004) « Présentation du volume. En quoi la phonologie est vraiment différente », *Corpus* 3. En ligne. URL : http://corpus.revues.org/193
Soum-Favaro, C., Coquillon, A. & Chevrot, J-P. (dir.) (2014) *La liaison : approches contemporaines*. Berne : Peter Lang.

Angus Grieve-Smith (Columbia University)
grvsmth@panix.com;

A representative theater of corpus for more accurate usage-based linguistics

The founding principle of linguistic theories like reduction, analogical extension and entrenchment (Bybee and Thompson 1997) is that the usage of today shapes the language of tomorrow. Collocations that are used frequently today will become shorter and simpler. Constructions that are used in a wide variety of contexts (type frequency) are more likely to be extended to new contexts than competing constructions. More frequent collocations tend to resist this extension.
To test these usage-based theories, we would ideally compare the language produced by an individual or individuals in a given period, with the language those individuals had previously been exposed to, both in comprehension and production. Even in this age of ubiquitous recording there is no complete record of either of those ideal corpora, for any individual. The best we can do is explicitly substitute samples that we believe to be close enough.
While the most common genre of language is spontaneous conversation, we do not have large, systematic or reliable samples of spontaneous conversation until recent decades. The best we can do is explicitly substitute genres that we believe to be close enough, such as theater and personal letters, keeping in mind that the language of both can be biased and artificial.
Not all historical texts have been preserved, made available in archives, and made available online. The best we can do is to explicitly acknowledge the missing texts, and identify systematic biases that operate in these processes.
Only a tiny fraction of all archived texts have made it into the canon, but many of our corpora are based on canonical texts. A notable example is FRANTEXT, created on the basis of a "principle of authority" (Imbs 1971) that chose texts based on their frequency of mention in literary histories. By text count, some authors are widely overrepresented: the corpus includes twelve plays by Théodore Leclerc, thirteen by Alfred de Musset and seven by Paul Claudel.
This is a problem that we can do something about, by sampling plays from an exhaustive catalog. One promising sampling frame is Wicks's list of all plays that premiered in public in Paris in the

nineteenth century, totaling over 30,000. The first volume (Wicks 1950), covering the years 1800-1815, contains over 3100 plays. I have created an initial corpus from a random sample of thirty-one plays (one percent) extracted from the first volume, and obtained copies of twenty-four of those plays.

The difference between the two corpora can be seen by brief investigations of well-known variables. For example, on average in the four theatrical texts for this period in FRANTEXT, 49% of negated declarative sentences used *ne* ... *pas* , 21% *ne* ... *point* , and 30% *ne* alone. In the fifteen plays currently available from the sampled corpus, we find on average 75% *ne* ... *pas* , 10% *ne* ... *point* and 15% *ne* alone ($p < 0.01$). 1.11% of non-interrogative sentences had a left-dislocated constituent in the sampled plays, compared to 0.40% in the FRANTEXT plays ($p < 0.005$).

Bibliography

Bybee, J. and S. Thompson. (1997). "Three frequency effects in syntax." *Proceedings of the Annual Meeting of the Berkeley Linguistics Society.*

FRANTEXT. (2017). Retrieved from http://www.cnrtl.fr/corpus/frantext/

Imbs, P. (1971). *Trésor de la langue française*. Paris : CNRS

Wicks, C.B. (1950,1953) *The Parisian Stage*. Tuscaloosa: University of Alabama