BERNHARD WÄLCHLI (Stockholm University)

Theory of similarity: three different kinds of "topic markers", a corpus approach to typology

Most linguistic theories heavily rely on notions of identity, non-identity and partial identity. This paper deals with linguistic typology, where examples of such notions are universals, language-particular categories and types (such as VO vs. OV word order). Identity and non-identity approaches entail considerable disagreement in assessing the amount of cross-linguistic diversity. They cannot appropriately capture the most likely scenario in which many linguistic categories are very similar, but hardly ever fully identical. This paper is a plea for a theory of similarity that still has to be developed.

As the name "typ"ology suggests, most typological investigations take for granted that languages fall into "types", a view that is hard to challenge when data is collected with high degree of data reduction. This paper pursues a corpus-linguistic approach to linguistic typology where exemplar data is collected in a massively parallel text (translations of the New Testament) and compared cross-linguistically on the level of individual examples.

The elements considered express various kinds of aboutness, including (i) "topic markers", such as Japanese *wa* and Quechua *-qa*, (ii) referentials or 'about'-expressions ("reported topics"), such as English *about*, and (iii) "European-style topic markers", such as English *as to* or French *quant à*. These phenomena have in common that they are under-researched in large-scale typological investigations, and (i) and (ii) are not usually considered to have much in common, as topics are dealt with in the context of information structure and referentials are rather viewed as semantic roles or adposition/case.

The paper addresses a number of notions that have played a major role in typology both from a methodological and from a theoretical point of view, notably (i) Haspelmath's comparative concepts (Lazard's *cadres conceptuelles arbitraires*), (ii) functional domains, (iii) gram types, (iv) implicational hierarchies, and (v) semantic maps/conceptual space. It is considered to what extent these notions are applicable to kinds of aboutness-markers both from a methodological and from a theoretical point of view when comparing languages on the level of exemplars, and also taking into account diachronic grammaticalization paths.

There is considerable disagreement in the literature about whether elements often referred to as "topic markers" really mark topics and if yes, what kinds of topics. It is not entirely clear to what extent these items have semantic, syntactic and/or pragmatic functions. They tend to have very high text frequency and highly specific distributions across texts, while there are at the same time few contexts where their use is strictly obligatory, which makes it difficult to identify prototypical functions. The investigation reveals that the markers considered very often cumulate the expression of different functions. All this suggests that cross-linguistic research is crucial for identifying general dimensions of variability. Cross-linguistic corpus research is a way for splitting up the multifunctionality of highly frequent language-specific markers. It is argued that multifunctional markers are better addressed within a framework of similarity than frameworks of identity or non-identity.

MARIA KHOLODILOVA, NATALIA LOGVINOVA, SOFIA OSKOLSKAYA (St. Petersburg University / Institute for Linguistic Studies, Russian Academy of Sciences)

Syntactically conditioned word-initial voicing in poshkart chuvash

The criteria used to delimit different levels of phonological structure have become subject of a few recent cross-linguistic studies concerned with (non-)convergence of these criteria. The main focus of these studies has been the level of phonological word[1], see, among others, (Schiering et al. 2010; Tallman 2020). The present study of word-initial voicing in the Poshkart dialect of Chuvash is designed to contribute to the discussion of the next phonological level suggested in Prosodic Phonology (Nespor, Vogel 1986/2007), namely phonological phrase. Sandhi phenomena, including juncture voicing, are among the most traditional criteria for phonological phrasehood. However, the study of sandhi demonstrates a bias towards qualitative data, which undermines the possibly gradual nature of these phenomena, cf. the scarcity of quantitative data in the paper collection edited by Andersen (1986).

The present study is based on quantitative data extracted from a corpus of spoken Poshkart Chuvash.

Chuvash obstruents undergo voicing in a position before a vowel if preceded by a vowel or a sonorant. With some notable exceptions partly described for the Poshkart dialect by Ašmarin (1898), this process is obligatory word-internally[2]. Crucially, voicing is also allowed under the same phonological conditions word-initially, resulting in structures like *petʲə-n **b**ørʦ-ë* (Petya-GEN house-P_3) ‘Petya’s house’, the stand-alone form of the second noun being ***p**ørʦ-ë*. Voicing never occurs word-finally even if the phonological conditions are met, cf. *torat / *tora**d** iltʦë* (branch took) ‘he took a branch’.

In all the known cases word-initial voicing is non-obligatory and shows a lot of variation in speech. Therefore, a study was conducted in order to determine the contexts which are most likely to trigger the voicing. Only the contexts which meet the phonological conditions were taken into account.

The study showed a few statistically significant correlations between the more syntactically cohesive structures and a higher frequency of voicing, including the following.
1) The predicates *pol* ‘be’ and *por* ‘exist’, unlike the other predicates, show an overwhelming tendency for voicing after subjects.
2) Postverbal NPs tend not to undergo voicing, while the juncture between a preverbal NP and the verb is a frequent position for voicing.
3) Postpositions following their complements undergo voicing more frequently than nouns following their modifiers.

Most crucially, different syntactic contexts don’t fall into just two groups according to the frequency of voicing. This suggests a model that allows varying degrees of phonological and syntactic integration along the lines of the quantitative studies on intonational contouring and pausation by Croft (1995) and Givón (1991) rather than an all-or-none approach of Prosodic Hierarchy (Nespor,

[1] See, however, a recent discussion in lingtyp mailing list “What’s the point of the phonological phrase?” (http://listserv.linguistlist.org/pipermail/lingtyp/, December 2019, and January 2020).

[2] Arguably, voicing in Poshkart Chuvash is phonological rather than phonetic, however, the discussion of this point is beyond the scope of this study.

Vogel 1986/2007). Moreover, the difference between the copula *pol* 'be' and other predicates suggests that the effect under consideration has access to lexical information, which is considered impossible for post-lexical processes in Prosodic Phonology.

References

Andersen H., ed. (1986). *Sandhi phenomena in the languages of Europe.* Berlin: Mouton de Gruyter.

Ašmarin N. I. (1898). *Materialy dlja izslědovanija čuvašskago jazyka.* Kazan: Tipo-litografija Imperatorskago Universiteta.

Croft, W. Intonation units and grammatical structure. *Linguistics* 33, 5, 1995. P. 839–882.

Givón, T. (1991). Some substantive issues concerning verb serialization: Grammatical vs. cognitive packaging. *Serial verbs: Grammatical, comparative and cognitive approaches*, ed. by C. Lefebvre. Amsterdam: Benjamins. P. 137–184.

Nespor, M. & I. Vogel. *Prosodic Phonology.* Berlin: Walter de Gruyter, 2007. — 1st ed. 1986.

Schiering, R., Bickel, B., & K. A. Hildebrandt (2010). The prosodic word is not universal, but emergent. *Journal of Linguistics* 46, 3. P. 657–709.

Tallman, Adam J. R. (2020). Beyond grammatical and phonological words. *Language and Linguistics Compass* 14, 2.

COSTAS GABRIELATOS (Edge Hill University, UK)

A corpus-based comparison of the explanatory power of linguistic theories: The case of the modal load in if-conditionals

Corpus-based examinations of the extent of modal marking (*modal load, ML*) in *if*-conditionals in the British National Corpus (BNC) have revealed that they have a significantly higher ML than average, as well as a higher ML than conditionals with other subordinators (e.g. *assuming*), conditional concessives, and non-conditional constructions with *if* (Gabrielatos 2007, 2010, 2019).

Explanations for these ML patterns are sought in the tenets of two recent linguistic theories: *Construction Grammar* (CxG) (e.g. Fillmore 1998) and *Lexical Grammar* (LG) (e.g. Sinclair 1996). The juxtaposition was motivated by the significant overlap in their tenets. Both theories take into account meaning (semantic and pragmatic), as well as lexical and grammatical factors. The central difference between the two theories is that LG gives clear prominence to lexis over grammar, whereas CxG posits no distinction between them. This study examines whether the ML of *if*-conditionals can be explained by recourse to the semantic preference of the word *if* (LG), or by recourse to the semantic component of conditional constructions (CxG).

The study uses eleven random samples of 1,000 *s-units*[1] from the written BNC:
- Written British English seen as a whole (*baseline*).
- Conditionals with other subordinators (*assuming, in case, on condition, provided, supposing, unless*).
- Conditional-concessives with *even if* and *whether*.
- Non-conditional structures taken collectively.
- Non-conditionals with subordinators containing *if*: indirect interrogatives with *if* and structures of comparison with a*s if*, as well as the same type of structures introduced by *whether* and *as though*, respectively.
- Non-conditionals introduced by the conjunctions *when* and *whenever,* as they have been compared to unmodalised conditionals (e.g., Athanasiadou & Dirven 1996:617, 1997:62; Palmer 1990:174-175).

The methodology combined manual annotation and quantitative analysis. The ML was established through two complementary metrics: *modal density* (MD) and *modalisation spread* (MS) (Gabrielatos 2010: 50-52). MD is the average number of modal markings per clause, and is expressed as the number of modal markings per 100 clauses. MS is the proportion of constructions that carry at least one modal marking, and is expressed as the percentage of modalised constructions. The size of similarities/differences in MD and MS values was also examined using hierarchical cluster analysis (Gabrielatos 2010: 52-54). The comparisons of MD and MS also take into account the statistical significance of differences, using the Bayesian Information Criterion (BIC) (Wilson, 2013).

The analysis provided strong indications that CxG rather than LG can best account for the emerging ML patterns. It was also shown that ML patterns are sensitive to different combinations of constructional attributes, as it would be predicted by the CxG principle of no synonymy (Goldberg 1995). This suggests that subordinators, rather than being the core of a lexical item (as LG would

[1] An s-unit is a stretch of text delimited on either side by a sentence-boundary marker (e.g. full-stop, question mark) (Sperberg- McQueen and Burnard 2007).

posit), are better seen as one of many components defining a construction. Consequently, if a semantic attraction of the subordinator can be posited, this has to be understood as being influenced by the type of construction that the subordinator is used in. In this light, semantic preference could be more usefully treated as part of the semantic component of a construction.

References

Athanasiadou, A. & Dirven, R. 1996. Typology of *if*-clauses. In Casad, E.H. (ed.), *Cognitive linguistics in the Redwoods: The expansion of a new paradigm in linguistics* (pp. 609-654). Berlin: Mouton de Gruyter.

Athanasiadou, A. & Dirven, R. 1997. Conditionality, hypotheticality, counterfactuality. In Athanasiadou, A. & Dirven, R. (eds.) *On Conditionals Again* (pp. 61–96). Amsterdam: John Benjamins.

Ball, C.N. 1994. Automated text analysis: Cautionary tales. *Literary and Linguistic Computing* 9(4), 265-302.

Croft, W. & Cruse, D.A. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Fillmore, C.J. 1998. The mechanisms of "Construction Grammar". In Axmaker, S. Jaisser, A. & Singmaster, H. (eds.) *General Session and Parasession on Grammaticalization.* Proceedings of the Fourteenth Annual Meeting of Berkeley Linguistics Society, February 13-15, 1998 (pp. 35-55). Berkeley: Berkeley Linguistics Society.

Gabrielatos, C. 2007. If-conditionals as modal colligations: A corpus-based investigation. In Davies, M., Rayson, P., Hunston, S. & Danielsson, P. (eds.) *Proceedings of the Corpus Linguistics Conference: Corpus Linguistics 2007*. Birmingham: University of Birmingham.

Gabrielatos, C. 2010. *A corpus-based examination of English if-conditionals through the lens of modality: Nature and types*. Unpublished PhD thesis, Lancaster University.

Gabrielatos, C. 2019. *If*-conditionals and modality: Frequency patterns and theoretical explanations. *Journal of English Linguistics*, 47(4), 301-334.

Goldberg, A. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.

Palmer, F.R. 1990. *Modality and the English Modals* (2nd ed.) Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Sinclair J.McH. 1996. The search for units of meaning. *Textus* 9(1), 75-106.

Sperberg-McQueen, C.M. & Burnard, L. 2007. *TEI P5: Guidelines for electronic text encoding and interchange*. The Text Encoding Initiative Consortium. Available online at http://www.tei-c.org/P5/Guidelines/AI.html

Wilson, A. 2013. Embracing Bayes factors for key item analysis in corpus linguistics. In Bieswanger, M. & Koll-Stobbe, A. (eds.) *New Approaches to the Study of Linguistic Variability* (pp. 3-11). Frankfurt: Peter Lang.

RANIA NAYEF AL-AQARBEH (Mu'tah University)

Resumption and Islands in Jordanian Arabic

This study investigated the amelioration effect of Resumption in *wh*-interrogatives with islands in spoken Jordanian Arabic (JA). The general assumption in previous theoretical and qualitative research is that resumption ameliorates structures with island violations in languages with grammatical resumption that allows Resumptive Pronouns (RPs) as a grammatical option, e.g. Arabic, and languages with intrusive resumption in which RPs are not a grammatical option, e.g. English (McCloskey, 2006) Most of the recent quantitative and experimental research has been devoted to languages with intrusive resumption concluding that the ameliorating effect of resumption if exists is very small and island constructions are rated nearly unacceptable with gaps and RPs even if the latter is insignificantly higher. Only a few experimental studies have addressed this issue in languages with grammatical resumption (Keshev & Meltzer-Asscher 2017) and mainly in visual mode of presentation. To that end, we report here a series of auditory judgment experiments on spoken Jordanian Arabic. We tested four island types (whether, wh-, complex NP, and adjunct islands) in wh-interrogative dependency using the factorial definition of island effects (Sprouse et al. 2012) to quantify the size of the island effects for both gaps and resumptive pronouns (and therefore the difference between them). We designed the experiments to be auditory experiments using a 7-point rating scale using PennController for IBEX (Drummond 2019, Zehr and Schwarz 2018). The factorial design has three factors: DEPENDENCY manipulating the location of the gap (matrix/embedded), STRUCTURE manipulating the structure of the critical clause (non-island/island), and RESUMPTION manipulating the tail (RP/gap). Due to the impossibility of an RP in the matrix subject position in Jordanian Arabic, the result is six conditions in a 2x2+2 design. Each participant completed an experiment that consisted of 54 items: 6 practice items, 24 experimental items, and 24 filler items pseudorandomized and distributed among experimental lists using a Latin square.

The results reveal statistically significant superadditive interactions which means that Jordanian Arabic demonstrates all four island types with wh-dependencies and gaps. They, though, unravel a superficially similar effect to amelioration by resumption, but the reason for the amelioration is not an increase in the acceptability of island/long (island-violating) conditions (the bottom right point), but rather a decrease in the acceptability of non-island/long sentences with resumption (the top right point). This suggests that resumption is dispreferred with non-island sentences and wh-dependencies in Jordanian Arabic. Our results suggest that resumption is dispreferred with wh-dependencies, and consequently does not result in any amelioration of island effects with wh-dependencies. This is contrary to general claim for resumption languages, and closer in behavior to the results found for intrusive resumption languages in the experimental literature.

These results support the suggestion from previous studies that the amelioration effect in grammaticalized resumption languages may either be smaller or more variable than previously thought. The scope of that variability is likely to lead to refinements to the theory of A'-dependencies and island effects.

References

Drummond, Alex. 2019. *Ibex Farm.* Available at: http://spellout.net/ibexfarm

Keshev, Maayan and Aya Meltzer-Asscher. 2017. Active dependency formation in islands: How grammatical resumption affects sentence processing. *Language,* 93, 549–68.

McCloskey, Jim. 2006. Resumption. In M. Everaert & H. van Riemsdijk (Eds.), *The Blackwell companion to syntax* (pp. 84–117). Blackwell.

Sprouse, Jon, Matthew Wagers, and Colin Phillips. 2012. A test of the relation between working memory capacity and syntactic island effects. *Language* 88: 82–123.

Zehr, Jérémy, and Florian Schwarz. 2018. PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832

Annex 1

Examples of the islands and dependencies tested

a.	Weiʃ sa'al Thaa'r itha is-siinama 9araD-at __? What asked Thaa'r whether the-cinema presented __	whether
b.	Weiʃ sa'al Bahaa' leiʃ Amaanii ʃaTab-at __? What asked Bahaa why Amani crashed __	wh
c.	Weiʃ naʃar Saamii il-iʃa9ah innu IntiSaar iʃtar-at __? What spread Sami the-rumor that Intisar bought __	complex np
d.	Weiʃ zi9il NaSer la'innu il-maktabih Taba9-at __? What got-angry Naser because the-press printed __	adjunct
e.	B-9rif il-mudiir illi il-skirteirah sa'al-at itha il-idarah ixtaar __. know the-manager who the-secretary asked whether the-board chose __.	whether
f.	ʃif-it il-asatitheh illi Sa7ab-ak sa'al leiʃ iT-Tullab wadda9-u __. saw the-teachers who friend-your asked why the-students said-farewell-to __.	wh
g.	B-9rif is-sikirteir-ah illi il-katib-ih simi9-it il-iʃa9ah innu il-mudiir itjawwaz __. know the-secretary who the clerk heard the rumor that the principal married __	complex np
h.	B-9rif il-mudiir illi il-binit inbasaT-at la'innu jaar-na 9azam __. know the-manager who the-girl felt-happy because neighbor-our invited __.	adjunct

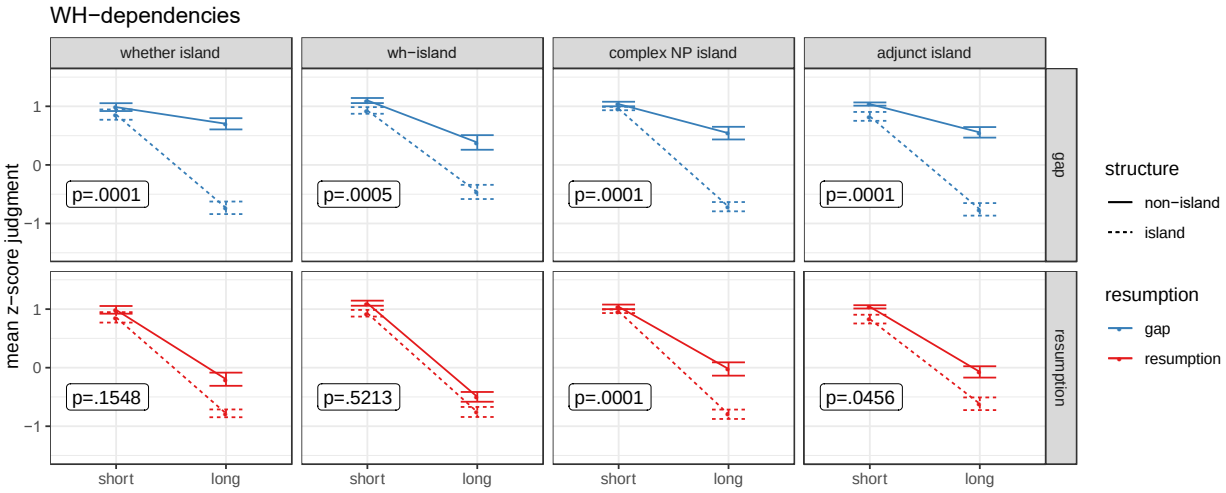

Figure 1: Results for WH-dependencies in Jordanian Arabic.

SASCHA DIWERSY (Université Paul-Valéry Montpellier III)

The company it keeps... Collocation, combinatorial profiles and the similarity of meaning

In recent years the so called distributional hypothesis introduced in the 1950s by Z. Harris as well as J. R. Firth (Harris 1954; Firth 1968 [1957]) gained considerable attention in the field of natural language processing. Although it has become commonplace since Mikolov et al.'s pioneering work (2013) to consider that vector representations based on cotextual data are best suited to identify the similarity of words, it is less obvious to conceptualise in a linguistically concise way what is meant by similarity in the context of these works. The aim of our talk is to give a theoretical and empirical assessment of this question, confronting at the same time the well established conceptual apparatus to account for similarity especially in the field of lexical semantics to the notion of meaning by collocation (Firth 1957).

References

Firth, John Rupert. (1957). Modes of meaning. In *Papers in Linguistics 1934-1951,* 190–215. London [e.a.]: Oxford University Press.

Firth, John Rupert. (1968). A Synopsis of Linguistic Theory, 1930-55. In Palmer, Frank Robert (ed.): *Selected Papers of J. R. Firth (1952-59),* 168–205. London [e.a.]: Longmans.

Harris, Zellig S. 1954. Distributional Structure. *Word* 10, 146–162.

Mikolov, Tomas, Chen, Kai, Corrado, Greg & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint* arXiv:1301.3781. [url: https://arxiv.org/pdf/1301.3781.pdf]

FEDERICA BEGHINI (University of Padova)

Conception et construction des corpus pour l'analyse stylistique intégrée d'une œuvre littéraire

Le projet en cours d'analyse stylométrique de l'œuvre de Milan Kundera, menée à l'aide des instruments textométriques, se base bien évidemment sur un corpus électronique. Nous voudrions discuter ici des critères de construction du corpus en vue des questions stylistiques auxquelles nous aimerions répondre.

L'objectif principal de notre étude est de détecter par contraste les éléments qui définissent le style de Kundera, en partant d'une comparaison de sa production avec les textes les plus significatifs des littératures française et tchèque de ces années-là. En particulier, les constantes et les variantes stylistiques seront examinées en diachronie, des années 60 aux années 2010 – en tenant compte des facteurs liés à l'âge de l'écrivain et à l'évolution de sa langue, à partir des traductions françaises de la langue source tchèque, jusqu'aux textes directement rédigés en français – et en synchronie – en étudiant les différences de contenu et de genre (narrative, essais). En outre, pour ce qui est des premiers textes de Kundera, l'interférence de la figure du traducteur sera examinée, dont le travail a ensuite été révisé par l'auteur lui-même.

Le défi du présent travail consiste donc à tester les outils textométriques pour voir si et à quel point ils peuvent aider à définir l'empreinte stylistique d'un auteur au profil linguistique très particulier. Pour ce faire, il est fondamental soient adaptées à l'objectif de la recherche tant les hypothèses qualitatives qui sous-tendent le processus de construction des corpus, que la définition du réseau d'analyses auxquelles ces derniers seront soumis.

Aussi notre communication se donnera-t-elle pour tâche de décrire minutieusement le processus de construction des corpus, avec ses enjeux, ses exigences et ses risques.

Pour cette raison, après avoir brièvement décrit le projet pour lequel le corpus a été établi, son objectif et ses méthodes, avec une attention particulière aux logiciels de textométrie choisis, nous exposerons les lignes directrices qui ont sous-tendu le processus de sélection des textes et les motivations à la base de leur regroupement en différents corpus : le corpus d'étude – l'*Œuvre I, II* de Kundera (Gallimard, Bibliothèque de la Pléiade) – et trois corpus de référence. Leur structure sera décrite en détail, ainsi que le procédé de résolution des choix problématiques. Nous présenterons aussi leur préparation pour le traitement informatisé, à savoir le nettoyage et le procédé de formatage appropriés à chaque logiciel. La lemmatisation et l'étiquetage morphosyntaxique seront pris en charge par les logiciels et seront préalables aux analyses présentées ci-dessus. En particulier, les éléments suivants feront l'objet de l'étude : la structure du vocabulaire ; la phrase et son rythme ; les aspects morphologiques et syntaxiques ; le contenu thématique.

Enfin, seront profilées les analyses comparatives que ces corpus rendent possibles : analyse endogène du corpus d'étude, études comparatives entre les sous-corpus d'étude et entre le corpus d'étude et les corpus de référence. Ces analyses endogènes et exogènes viseront à faire ressortir, par le biais de diverses comparaisons et contrastes, le noyau prototype du style de l'écrivain.

Bibliographie

Méthodes d'analyse linguistique et statistique du corpus

Biber D., Conrad S., Reppen R. (1998), *Corpus linguistics, Investigating Language, Structure and Use*, Cambridge, Cambridge Approaches to Linguistics.

Brunet, É. (1981), *Le vocabulaire français de 1789 à nos jours*, Paris – Genève, Honoré Champion – Slatkine.

Brunet, É. (2000), « Qui lemmatise dilemme attise » in *Scolia, 11èmes rencontres linguistiques en Pays Rhénan*, José L. et Theisse A. (éds.), Strasbourg, Publications de l'Université Marc Bloch, n°13.

Brunet, É. (2009), *Comptes d'auteurs*, Paris, Honoré Champion.

Brunet, É. (2011), *Ce qui compte – Ecrits choisis, tome II*, Paris, Honoré Champion.

Brunet, É. (2016), *Tous comptes faits. Écrits choisis*, tome III : Questions linguistiques, Bénédicte Pincemin (éd.), Paris, Honoré Champion, coll. « Lettres numériques ».

Cortelazzo M. *et al.* (2012), "Una versione iterativa della distanza intertestuale applicata a un corpus di opere della letteratura italiana contemporanea", *JADT 2012 Actes des 11es Journées internationales d'analyse statistique des données textuelles*, p. 295-307.

Cortelazzo, M.A. (2013a), "Metodi qualitativi e quantitativi di analisi dei testi", *Contemporanea*, 2, p. 299-310.

Cortelazzo, M.A., Nadalutti, P., Tuzzi, A. (2013b), "Improving Labbé's Intertextual Distance: Testing a Revised version on a Large Corpus of Italian Literature", *Journal of Quantitative Linguistics*, 20(2), p. 125-152.

De Mauro T., Chiari I. (2005), *Parole e numeri – Analisi quantitative dei fatti di lingua*, Aracne, Roma.

Dodge, Y. (éd.), (2003), *The Oxford Dictionary of Statistical Terms*, Oxford, Oxford University Press.

Favretti, R. R. (2000), "Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS", in R. Rossini Favretti (ed.), *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*, Bulzoni, Roma, p. 39-56.

Guiraud, P. (1960), *Problèmes et méthodes de la statistique linguistique*, Paris, Presses universitaires de France.

Holmes, D.I. (1998), "The evolution of stylometry in humanities scholarship, *Literary and Linguistic Computing*, 13(3), p. 111-117.

Kastberg Sjöblom M. (2002), « L'écriture de J.M.G Le Clézio, une approche lexicométrique », Thèse de doctorat, Université de Nice – Sophia Antipolis. Publication électronique dans *Texto ! Textes et Cultures*, revue électronique de sémantique des textes, F. Rastier (éd.), Paris, l'Institut Ferdinand de Saussure, Maison des Sciences de l'Homme.

Kastberg Sjöblom M. (2002), « Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus », in *JADT 2000, 6èmes Journées internationales d'Analyse statistique des Données Textuelle*, Morin A., Sébillot P. (éds.), Irisa, Inria, Saint-Malo, 13-15 mars, p. 391-402.

Kastberg Sjöblom M. (2004), « Comment l'ordinateur peut-il servir dans l'étude stylistique d'un texte littéraire et de quelle façon l'analyse de la distribution des parties du discours peut-elle contribuer à la compréhension des textes ? », M. Ballabriga et F.-C. Gaudard éds.), *Champs du Signe*, Toulouse, Editions Universitaires du Sud, p. 119-152.

Labbé C., Labbé D. (2005), "A Tool for Literary Studies: Intertextual Distance and Tree Classification", *Literary & Linguistic Computing*, 2006, 21(3), p. 311-326.

Labbé C., Labbé, D. (2001), "Inter-Textual Distance and Authorship Attribution. Corneille and Molière", *Journal of Quantitative Linguistics*, 8(3), p. 213-231.

Lebart, L., Pincemin, B., Poudat, C. (2019), *Analyse des données textuelles*, Québec, Presses de l'Université du Québec.
Lebart, L., Salem, A. (1994), *Statistique textuelle*, Paris, Dunod.
Magri-Mourgues, V. (2010), « Stylistique et statistiques. Le corpus textuel et hyperbase », In : *Stylistique ?*, Rennes : Presses universitaires de Rennes.
Magri-Mourgues, V. (2011), « Analyse textométrique et interprétation littéraire – Hyperbase, Rousseau et les Lumières », *Travaux neuchâtelois de linguistique* (TRANEL), n°5, p. 77-93.
Picard J., Pibarot A. et Labbé D. (1995), « Un outil de statistique textuelle : le lemmatiseur », in Travaux scientifiques C.R.S.S.A., n° 16, p. 395-396.
Pincemin, B. (2012), « Sémantique interprétative et textométrie », *Texto ! Textes et Cultures*, 17(3), http://www.revue-texto.net/index.php?id=3049, consulté le 26 avril 2020.
Pincemin, B. (2018), « Sept logiciels de textométrie », Archives ouverts HAL, CNRS/CCSD, Villeurbanne, https://halshs.archives-ouvertes.fr/halshs01843695, consulté le 26 avril 2020.
Pincemin, Bénédicte (2020), « La textométrie en question », *Le Français Moderne – Revue de linguistique Française*, CILF (conseil international de la langue française).
Rastier, Fr. (2011), *La mesure et le grain. Sémantique de corpus*, Paris, Honoré Champion, Collection Lettres numériques.
Sjöblom M. K. (2006)*, L'écriture de J.M.G. Le Clézio. Des mots aux thèmes*. Paris, Honoré Champion.
Tuzzi A., Cortelazzo M. A. (2018a), "What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer", *Digital Scholarship in the Humanities*, 33(3), p. 685–702.
Tuzzi A., Cortelazzo, M. (dir.; 2018b), *Drawing Elena Ferrante's Profile*, Padova, Padova University Press.

Ouvrages généraux – linguistique, stylistique, littérature, traduction

Adam, J.-M. (1989), *Le texte descriptif*, Paris, Nathan.
Adam, J.-M. (1992), *Les textes, types et prototypes. Récit, description, argumentation, explication et dialogue*, Paris, Nathan, 2001.
Calas, F. (2007), *Introduction à la stylistique*, Paris, Hachette Supérieur.
Fromilhague C., Sancier-Château A. (2002), *Introduction à l'analyse stylistique*, Paris, Armand Colin.
Herschberg Pierrot, A. (2003), *Stylistique de la prose*, Paris, Éditions Belin.
Histoire de la littérature française du XXe siècle : Tome 2 – après 1940 (dir. M. Touret), Rennes, Presses universitaires de Rennes.
Holý, J. (2008), *Writers Unders Siege – Czech Literature since 1945*, Sussex Brighton, Academic Press.
La Littérature française : dynamique et histoire, II (dir. J.-Y. Tadié), 2007, Gallimard.
Molinier, G. (1989), *La stylistique*, Paris, Presses Universitaires de France, 1991.
Narteau C., Nouailhac I. (2010), *Littérature française. Les grands mouvements littéraires du Moyen Âge au XX Siècle*, Paris, Flammarion.
Richter, M., Capatti A. (2000), *Meridiennes II. Histoire et anthologie de la Littérature Française*, Milano, Mondadori Bruno Scolastica.
Rybicki, J. (2012), "The great mystery of the (almost) invisible translator", *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*, a cura di Oakes, M. P. and Meng Ji, John Benjamins, Amsterdam, p. 231-248.
Venuti, L. (1992), *Rethinking translation: discourse, subjectivity ideology*, London – New York, Routledge.
Venuti, L. (1995), *The translator's invisibility: a history of translation*, London, Routledge.

Venuti, L. (1998), *The Scandals of translation towards an ethics of difference*, London – New York, Routledge.
Venuti, L. (2013), *Translation changes everything: Theory and Practice*, London – New York, Routledge.
Voisine-Jechova, H. (2001), *Histoire de la littérature tchèque*, Paris, Librairie Arthème Fayard.

Ouvrages de Milan Kundera

Kundera, M. (1968), *La plaisanterie*, Paris, Gallimard, coll. Du monde entier.
Kundera, M. (1970), *Risibles amours*, Paris, Gallimard, coll. Du monde entier.
Kundera, M. (1973), *La vie est ailleurs*, Paris, Gallimard, coll. Du monde entier.
Kundera, M. (1976), *La valse aux adieux*, Paris, Gallimard, coll. Du monde entier.
Kundera, M. (1979), *Le livre du rire et de l'oubli*, Paris, Gallimard, coll. Du monde entier.
Kundera, M. (1980), *La plaisanterie*, Paris, Gallimard, coll. Du monde entier.
Kundera, M. (1984), *L'insoutenable légèreté de l'être*, Paris, Gallimard, coll. Du monde entier.
Kundera, M. (2011), *Œuvre I et II*, Paris, Gallimard, Coll. Bibliothèque de la Pléiade, 2017.
Kundera, Milan (1985), *La plaisanterie*, Paris, Gallimard, Coll. Du monde entier, 1993.

Ouvrages sur Milan Kundera

Banerjee, M. N. (1990), *Terminal Paradox*, England, Faber and Faber, 1991.
Chvatik, K. (1995), *Le monde romanesque de Milan Kundera*, Paris, Gallimard, coll. Arcades.
Computers and the Humanities, 35(2), p. 193-214, disponible sur jstor.org.
Le Grand, E. (1995), *Kundera ou la mémoire du désir*, Paris, L'Harmattan.
Merrill T. C., (2013), *The Book of Imitation and Desire: Reading Milan Kundera with Rene Girard*, London, Bloomsbury Academy, 2014.
Misurella, F. (1990), *Understanding Milan Kundera: Public Events, Private Affairs*, New York, Grove Weidenfeld.
Mravlja, K. (2015), *Milan Kundera : entre l'original et la traduction – Comparaison des traductions françaises du roman tchèque La Plaisanterie*, Ljubljana, Univerza V Ljubljani.
Ricard, F. (2011), *L'ultimo pomeriggio di Agnes: saggio sull'opera di Milan Kundera* (traduzione di Paola Vallerga), Udine, Mimesis.
Rizek, M. (2001), *Comment devient-on Kundera ?*, Paris, L'Harmattan.
Thirouin, M.-O. et Boyer-Weinmann M. (éd.) (2009), *Désaccords parfaits – La réception paradoxale de l'œuvre de Milan Kundera*, Grenoble, Ellug.
Woods, M. (2006), *Translating Milan Kundera*, Clevedon, Multilingual Matters.

GUILLAUME GUEX (University of Lausanne)

Diachronic Word Embedding Methods: a comparative study for digital humanities and computational linguistic researchers.

Word embedding methods, also called *distributional representation* methods, have received considerable attention from the Digital Humanities and Computational Linguistics communities in the past few years, particularly since the publication of the famous *word2vec* method in 2013 (Mikolov et al., 2013), although the idea behind these methods is not new and can be traced back to the creation of *latent semantic analysis* (Deerwester et al., 1990). By using the *contexts* of words found in a given corpus, these methods automatically create low dimensional vectors representing each word appearing in the corpus and they place these vectors relatively to each other to construct a kind of *semantic map*: semantically close words lie in the same neighborhood (e.g. *cup* and *mug*) and some relationships are expressed geometrically (e.g. *king – man + woman ≈ queen*). The popularity of *word2vec*, as well as its subsequent and close competitors, such as *GloVe* (Pennington et al., 2014), *fastText* (Bojanowski et al., 2017) and *ELMo* (Peters et al., 2018), is not surprising, as the produced word vectors where used in a broad range of Natural Language Processing tasks with competitive results.

One of the main advantages of word embedding methods is that they produce *task agnostic results*, word vectors only reflect semantic relationships between words found in the corpus and thus can be used an exploratory tool. Therefore, an increasing number of researchers in various fields began to use them to extract semantic information from corpora in order to test different linguistical, cultural or historical hypotheses (see, e.g., Grayson et al., 2016, Kerr, 2017, Heuser, 2017). This enthusiasm was further increased as diachronic word embedding methods began to emerge, enabling to compute semantic trajectories of words regarding time (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016; Bamler & Mandt, 2017; Yao et al., 2018; Rudolph & Blei, 2018; Rosenfeld & Erk, 2018; Lauretig, 2019). These diachronic approaches allowed researchers to postulate laws of semantic changes for words (Dubossarsky et al., 2015; Hamilton et al., 2016), identify lexical replacement (Szymanski, 2017), predict periods of political insecurities (Mueller & Rauh, 2018) or find historical periods with high linguistic transformations (Jo & Algee-Hewitt, 2018), among others.

Nevertheless, this fast-developing field also received its share of criticisms (Dubossarsky et al., 2017; Hellrich & Hahn, 2016, 2017). The lack of gold standards, the difficulty to judge the quality of embeddings or to know which type of semantical relationships they capture (word similarity, association or relatedness, see Hill et al., 2015) and the lack of reproducibility of some methods (because of their random nature) were pointed out. Moreover, as the field is quite recent, researchers lack systematic and intensive comparison survey for available methods.

This presentation will come back to the most promising methods for diachronic word embeddings as well as derived time series for semantic studies. A complete comparative study concerning reliability, accuracy and limits of methods will be given, allowing Digital Humanities or Computation Linguistics researchers to select the most appropriate one regarding their data. This theoretical framework will be supported by case studies build on the French newspapers data from the *impresso project*[1].

[1] https://impresso-project.ch/

References

Bamler, R., & Mandt, S. (2017). Dynamic Word Embeddings. *arXiv:1702.08359 [cs, stat]*. http://arxiv.org/abs/1702.08359

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *ArXiv:1607.04606 [Cs]*. http://arxiv.org/abs/1607.04606

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Dubossarsky, H., Tsvetkov, Y., Dyer, C., & Grossman, E. (2015). A bottom up approach to category mapping and meaning change. *NetWordS*, 66-70.

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta Control : Laws of Semantic Change and Inherent Biases in Word Representation Models. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1136–1145. https://doi.org/10.18653/v1/D17-1118

Grayson, S., Mulvany, M., Wade, K., Meaney, G., & Greene, D. (2016, septembre 21). *Novel2Vec : Characterising 19th Century Fiction via Word Embeddings*. 24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), University College Dublin, Dublin, Ireland, 20-21 September 2016. https://researchrepository.ucd.ie/handle/10197/8360

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *arXiv:1605.09096 [cs]*. http://arxiv.org/abs/1605.09096

Hellrich, J., & Hahn, U. (2017). Don't Get Fooled by Word Embeddings-Better Watch their Neighborhood. *DH*.

Hellrich, J., & Hahn, U. (2016). Bad Company—Neighborhoods in Neural Embedding Spaces Considered Harmful. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2785–2796. https://www.aclweb.org/anthology/C16-1262

Heuser, R. J. (2017). Word Vectors in the Eighteenth Century. *DH*, 5.

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999 : Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, *41*(4), 665-695. https://doi.org/10.1162/COLI_a_00237

Jo, E. S., & Algee-Hewitt, M. (2018). The Long Arc of History : Neural Network Approaches to Diachronic Linguistic Change. *Journal of the Japanese Association for Digital Humanities*, *3*(1), 1-32. https://doi.org/10.17928/jjadh.3.1_1

Kerr, S. J. (2017). When Computer Science Met Austen and Edgeworth. *NPPSH Reflections*, *1*, 38-52.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal Analysis of Language through Neural Language Models. *arXiv:1405.3515 [cs]*. http://arxiv.org/abs/1405.3515

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically Significant Detection of Linguistic Change. *Proceedings of the 24th International Conference on World Wide Web*, 625–635. https://doi.org/10.1145/2736277.2741627

Lauretig, A. M. (2019). Identification, Interpretability, and Bayesian Word Embeddings. *arXiv:1904.01628 [cs, stat]*. http://arxiv.org/abs/1904.01628

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. http://arxiv.org/abs/1301.3781

Mueller, H., & Rauh, C. (2018). Reading Between the Lines : Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, *112*(2), 358-375. https://doi.org/10.1017/S0003055417000570

Pennington, J., Socher, R., & Manning, C. (2014). Glove : Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365 [cs]*. http://arxiv.org/abs/1802.05365

Rosenfeld, A., & Erk, K. (2018). Deep Neural Models of Semantic Shift. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 474–484. https://doi.org/10.18653/v1/N18-1044

Rudolph, M., & Blei, D. (2018). Dynamic Embeddings for Language Evolution. *Proceedings of the 2018 World Wide Web Conference*, 1003–1011. https://doi.org/10.1145/3178876.3185999

Szymanski, T. (2017). Temporal Word Analogies : Identifying Lexical Replacement with Diachronic Word Embeddings. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 448–453. https://doi.org/10.18653/v1/P17-2071

Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic Word Embeddings for Evolving Semantic Discovery. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*, 673-681. https://doi.org/10.1145/3159652.3159703

LUDOVIC MONCLA, DENIS VIGIER, KATHERINE MCDONOUGH, ALICE BRENON ET THIERRY JOLIVEAU (INSA Lyon/ Université Lyon 2/ The Alan Turing Institute and Queen Mary/ Université de Saint-Etienne)

Combinaison d'approches qualitative et quantitative pour le repérage et la classification des entités nommées dans l' Encyclopédie de Diderot et d'Alembert (1751-1772)

Dans cette communication, nous décrivons une méthodologie pour l'amélioration de la reconnaissance et de la classification des entités nommées (EN) (Chinchor & Marsh, 1998) pour des textes anciens de type encyclopédique (XVIII e siècle). La nature de ces documents pose de nombreuses limitations pour les approches existantes : i) l'état de la langue (orthographe, lexique, syntaxe), ii) une graphie non stabilisée des noms de pays et de régions éloignées de la France et a fortiori de l'Europe, iii) un discours géographique qui mêle descriptions de lieux réels aussi bien que mythologiques, de lieux évoqués seulement dans les Écritures ou encore de lieux mentionnés par des auteurs antiques mais dont l'existence est problématique voire exclue.

Les méthodes de reconnaissance et de classification des EN sont réparties en trois familles (Sekine & Eriguchi, 2000 ; Poibeau 2001) : symbolique, statistique et hybride. Les approches symboliques mettent en oeuvre des règles mises au point par des experts et mobilisent des ressources linguistiques (dictionnaires, lexiques, données d'ordre typographique, collocationnelles, ...). Les approches statistiques utilisent des méthodes d'apprentissage automatique qui nécessitent le plus souvent de grands corpus annotés. Elles construisent des modèles (arbres de décision, modèles probabilistes, ...) de manière supervisée mais dont les résultats obtenus, les règles apprises, ou les décisions prises, peuvent être complexes à comprendre. Enfin, les systèmes hybrides combinent méthodes symboliques et statistiques le plus souvent adaptées pour des méthodes semi-automatiques avec des interactions entre un opérateur et les algorithmes.

Une des originalités de notre approche consiste à combiner étroitement approches qualitative et quantitative. En premier lieu, par l'exploitation de certaines spécificités discursives propres au genre encyclopédique (Mazière, 1982; Cernuschi, 2018). Une telle approche rompt quelque peu avec celles traditionnellement mises en oeuvre en TAL (McDonough et al., 2019). En effet, selon Nadeau & Sekine, "The impact of textual genre (journalistic, scientific, informal, etc.) and domain (gardening, sports, business, etc.) has been rather neglected in the NERC literature" (Nadeau & Sekine, 2007). Ensuite, l'identification d'indices linguistiques de surface (McDonald, 1996) par une analyse textométrique du corpus au moyen de la plateforme TXM1 (Heiden, 2010). Ces indices peuvent être "forts" ou "faibles" et permettent d'aboutir à la catégorisation et à la sous-catégorisation des EN. Enfin, l'élaboration et l'implémentation de règles pour le repérage automatique de ces indices dans un système d'annotation. Pour cela, nous nous appuyons sur la plateforme existante PERDIDO2 (Moncla & Gaio, 2018) permettant l'annotation automatique d'information géographiques (noms de lieux, relations, spatiales, …) dans des textes. Nous proposons également une méthode permettant la combinaison et la pondération de ces indices "forts" et "faibles" ainsi qu'une évaluation sur l' *Encyclopédie* de leur apport pour les tâches de reconnaissance et de classification de EN.

Références

A. Cernuschi (2018) Une espèce d'ouvrage cosmopolite. Variations énonciatives dans les articles encyclopédiques des Lumières, de Chambers à l'Encyclopédie, Travaux Neuchâtelois de Linguistique, 69, 39-59.

N. Chinchor, E. Marsh (1998) MUC-7 information extraction task definition (version 5.1), Proceedings of MUC Vol. 7

S. Heiden (2010) The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation - PACLIC24 (p. 389-398).

Mazière, F. (1982), Cellule : un discours de vulgarisation dans les dictionnaires encyclopédiques, Langue française, 53, 62-77.

D. Nadeau, S. Sekine (2007) A survey of named entity recognition and classification, Lingvisticae Investigationes 30, 3

D.D. McDonald (1996) Internal and external evidence in the identification and semantic categorization of proper names, Corpus processing for lexical acquisition

K. McDonough, L. Moncla, M. van de Camp (2019) Named entity recognition goes to old regime france : geographic text analysis for early modern french corpora, International Journal of Geographical Information Science (IJGIS) 0, 1

L. Moncla, M. Gaio (2018) Services web pour l'annotation sémantique d'informations patiales à partir de corpus textuels, Revue Internationale de Géomatique (RIG) 28, 439

T. Poibeau (2005) Sur le statut référentiel des entités nommées, in Conférence Traitement Automatique des Langues 2005 (Association pour le Traitement Automatique des LAngues/LIMSI), pp. 173–183

S. Sekine, Y. Eriguchi (2000) Japanese named entity extraction evaluation: analysis of results, Proceedings of the 18th conference on Computational linguistics-Volume 2 (ACL), pp. 1106–1110

BARBARA MCGILLIVRAY (University of Cambridge and The Alan Turing Institute)

Modelling lexical semantic change using quantitative approaches: from Ancient Greek to emoji

Lexical semantic change, i.e. the phenomenon in which the semantics of lexical items changes over time, has been the object of qualitative research for over a century. Anthropological studies in linguistics (Boas 1911; Sapir 1912; 1928) and in conceptual history (Williams 1976; Richter 1995) have recognised the importance of this research to reach a better understanding of the dynamics of cultural, social and political systems. Philological methods (e.g., Kenny 1995, Wierzbicka 1997) and theoretical linguistics research (Geeraerts 2010; Koch 2016; Grondelaers et al. 2007) have also engaged with the analysis of language-internal aspects of this phenomenon.

Recent digitization efforts have now made it possible to access and mine large-scale digital collections of historical texts using automatic methods and investigate the question of semantic change over centuries at a new level of resolution. Easy access to very large born-digital collections from the web also allows us to study changes in contemporary language data spanning short time periods. What are the benefits and the limitations of applying large-scale quantitative methods to the study of such a complex phenomenon?

In this talk I will present my research on developing models for semantic change drawing on state-of-the-art quantitative methods relying on distributional semantics principles. I will share my experience of working at different scales and in a range of interdisciplinary projects, from Ancient Greek and Latin to Charles Darwin's letters, web archives, Twitter and emoji.

References

Boas, F. (1911). "Introduction", in Boas, F. (ed.), *Handbook* of *American Indian Languages,* Washington D.C.: Bureau of American Ethnology, Bulletin 40 (I), pp. 59-73.

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford: Oxford University Press

Grondelaers, Stefan, Speelman, Dirk and Geeraerts, Dirk (2007). Lexical variation and change. In *The Oxford handbook of cognitive linguistics*.

Kenny, Neil. (1995). Interpreting Concepts after the Linguistic Turn: The Example of curiosité in Le Bonheur des sages / Le Malheur des curieux by Du Souhait (1600)', in *Interpréter le seizième siècle*, ed. by John O'Brien (Michigan Romance Studies, XV, 1996), 241–70.

Koch, P. (2016). Meaning change and semantic shifts. In Päivi Juvonen and Maria Koptjevskaja Tamm (eds.), *The Lexical Typology of Semantic Shifts*, pages 21–66. De Gruyter Mouton, Berlin/Boston.

Richter, M. (1995). *The History of Political and Social Concepts: A Critical Introduction.* New York and Oxford: Oxford University Press.

Sapir, E. (1912). "Language and Environment", American Anthropologist 14, pp. 226-242.

Sapir, E. (1928). Proceedings, First Colloquium on Personality Investigation; Held under the Auspices of the American Psychiatric Association, Committee on Relations with the Social Sciences, New York: Lord Baltimore Press, pp. 77-80.

Wierzbicka, Anna. (1997). *Understanding Cultures Through Their Key Words: English, Russian, Polish, German, and Japanese.* Oxford: Oxford University Press.

Williams, Raymond. (1976). *Keywords. A Vocabulary of Culture and Society*. London: Fourth Estate.

ESSA ALI BATEL (University of Arizona)

Formulaic Expressions Processing in the First and Second Language

The presented paper presents a result of a study that examined advanced English non-native speakers' (L2) processing of formulaic expressions embedded within sentence contexts. According to Laufer & Waldman (2011), formulaic language is a combination of words that are of restricted co-occurrence in that individual words in these combinations are, for the most part, easily replaceable following the rules of grammar, yet not in all cases (e.g. *Tea* collocates with *strong* but not with *powerful*). These multi-word units are presumed to be stored in and retrieved from long-term memory as if they were single lexical units (Read & Nation, 2004; Wray & Perkins, 2000). In this sense and based on the Spreading Activation model (Collins & Loftus, 1975) and the Associative Relatedness model (Postman & Keppel, 2014), the recognition time (RT) of the noun "*tea*" after the incongruent adjective "*powerful*" should take longer compared to the RT of the same word after the congruent adjective "*strong*" due to the discrepancy between the predicted word and the printed word. Therefore, the incongruent preceding context has an inhibitory effect on the target word leading to a longer RT compared to the congruent context. In light of this research, this study sought to determine whether late L2 speakers of English show the same inhibitory effect as L1 speakers when a synonymous but not congruent word (i.e. *powerful*) preceded the target word in a formulaic sequence combination. The goal was to test whether in L2 learners, formulaic expressions are stored in and retrieved from long-term memory as if they were single lexical units, as they are in native speakers. The hypothesis was that L2 formulaic expressions are not stored in and retrieved from long-term memory as the case in L1 formulaic expressions.

To test this hypothesis, 24 adult advanced-level learners of English were recruited as the experiment group and 24 adult native speakers of English. The experiment's participants performed a Self-paced Reading task that consisted of 24 English sentences containing formulaic combinations (the sentences were divided into equal halves with each half containing either congruent or incongruent combinations). On a computer screen, each sentence was presented one word at a time and participants had to press the SPACE key to move from one word to the next. Participants were then prompted to answer a comprehension question at the end of each sentence. The RTs on critical words (i.e. the word *tea*) were compared in both congruency conditions. The results showed a significant shorter RT for the congruent combination compared to the incongruent combination in L1 participant, but not for L2 participants.

These results indicate that L2 learners do not attend to the same level of sensitivity to violations in multi-word formulaic sequences when one word in the formulaic combination is replaced by a close synonym. The findings also suggest that formulaic expressions are not stored in and retrieved from long-term memory for L2 learners as if they were single lexical units, at least not during online processing that simulates subconscious recognition of lexical items.

Bibliography

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. Language learning, 61(2), 647-672.

Postman, L., & Keppel, G. (Eds.). (2014). *Norms of Word Association*. Academic Press.

Read, J., & Nation, P. (2004). Measurement of formulaic sequences. *Formulaic Sequences: Acquisition, Processing and Use*, 23-35.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*(1), 1-28.

ELENI KARANTZOLA & YANNIS KOSTOPOULOS (University of the Aegean)

Reformulations as a diagnostic tool for corpora representativity.

In the last four decades, reformulation relations and linguistic items indicating reformulating operations have been extensively explored through a considerable number of theoretical perspectives (Fuchs 1982, Gülich & Kotschi 1983, Roulet 1987, Mann & Thompson 1988, Rossari 1990, Blakemore 1993, 2007, Burton-Roberts 1994, Bhagat & Hovy 2013) and in a significant amount of languages (Fløttum 1994, Rossari 1994, Vassiliadou 2004, Cuenca & Bach 2007, Del Saz 2007). Regardless of any remaining areas of controversy - especially on issues pertaining to reformulations taxonomy and reformulation markers description (see Murillo 2004, Steuckardt 2009, Vassiliadou in press) - previous investigations contributed important theoretical and empirical findings which largely clarified the notion of reformulation, and shaped the methodological tools for approaching reformulating operations. The application of these findings in discourse analysis revealed strong associations between certain reformulating operations and particular text-types (e.g. equivalence/paraphrastic reformulations and scientific discourse, see Thoiron & Béjoint 1991).

Adapting previous research on reformulation in a corpus-based approach to historical linguistics, in this paper we try to explore whether reformulation relations and reformulation markers can be used as diagnostic tools in historical corpora investigations. Precisely, we consider the possibility of treating linguistically marked reformulating operations as evidence of the representativity of a corpus, and as indications for locating a given text in the orality/literacy and the standard/vernacular continua. Our study focuses on Early Modern Greek (EMG, 16^{th}-17^{th} century), an understudied variety characterised by extended linguistic variation and absence of standardisation (Karantzola 2006). In what concerns the expression of reformulation, EMG exhibits at least five multifunctional reformulation markers, with only one of them surviving in contemporary Modern Greek.

Using data from a 155.000-words, genre-classified, anthology of texts (Kakoulidou-Panou, Karantzola & Tiktopoulou, in press), and adopting a quantitative-qualitative approach, we explore the associations between reformulation operations and markers, on the one hand, and text genre, on the other. According to our hypothesis, the existence of systematic, significant correlations of this sort, would suggest that reformulation markers and operations are indicative of corpora genre representativity. Furthermore, in our study we scrutinize the frequencies and functions of EMG reformulation markers, considering how reformulation markers can provide a diagnostic tool for identifying vernacular / non-vernacular tokens in historical corpora.

References

Bhagat, R. & E. Hovy. 2013. “What is a paraphrase?” *Computational Linguistics* 39(3): 463-472.
Blakemore, D. 1993. “The relevance of reformulations”. *Language and Literature* 2(2): 101-120.
Blakemore, D. 2007. “‘Or’-parentheticals, ‘that is’-parentheticals and the pragmatics of reformulation”. *Journal of Linguistics* 43: 311-339.
Burton-Roberts, N. 1994. “Apposition”. In Asher, R.E. & J.M.Y. Simpson (eds.) *The Encyclopaedia of Language and Linguistics Vol. I.* Oxford: Pergamon Press, 184-187.

Cuenca, M.J. & C. Bach. 2007. "Contrasting the form and use of reformulation markers". *Discourse Studies* 9(2): 149-175.
Del Saz Rubio, M.M. 2007. *English Discourse Markers of Reformulation: A Classification and Description.* Bern: Peter Lang.
Flottum, K. 1994. "À propos de c'est-à-dire et ses correspondants norvégiens". *Cahiers de linguistique française 15: 109-130.*
Fuchs, C. 1982. *La paraphrase.* Paris: PUF.
Gülich, E. & T. Kotschi. 1983. "Les marqueus de la reformulation paraphrastique". *Cahiers de linguistique française* 5: 305-351.
Kakoulidi-Panou, E., Karantzola, E., & K. Tiktopoulou (in press). *Demotic Prose Texts of the 16th c.* [in Greek]. Thessaloniki and Athens: Centre for Greek Language and MIET.
Karantzola, E. 2006. "Sur la langue 'populaire' de la Grèce ottomane et vénitienne". *Revue des études néohelléniques* 2 (n.s.): 107-122.
Mann, W.C. & S. A. Thompson. 1988. "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization". *Text* 8(3): 243-281.
Murillo, S. 2004. "A relevance reassessment of reformulation markers". *Journal of Pragmatics* 36: 2059-2068.
Rossari, C. 1990. "Projet pour une typologie des opérations de reformulation". *Cahiers de linguistique française* 11: 345-359.
Rossari, C. 1994. *Les opérations de reformulation*. Berne: Peter Lang.
Roulet, E. 1987. "Complétude interactive et connecteurs reformulatifs". *Cahiers de linguistique française* 8: 111-140.
Steuckardt, A. 2009. "Décrire la reformulation: le paramètre rhétorique". *Cahiers de praxématique* 52: 159-172.
Thoiron, P. & H. Béjoint. 1991. "La place de la reformulation dans les textes scientifiques". *Meta* 36(1): 101-110.
Vassiliadou, H. (in press). "Peut-on aborder la notion de reformulation autrement que par la typologie de ses marques? Pour une analyse *s*émasiologique et onomasiologique". In Inkova, O. (ed.), *Autour de la reformulation*. Geneva: Droz.
Vassiliadou, H. 2004. *Les connecteurs c'est-à-dire (que) en français et ðilaði en grec. Analyse syntaxique et sémantico-pragmatique.* Thèse de Doctorat, Strasbourg.

DARIO COMPAGNO (Université de Lorraine)

Is a Finalistic Interpretation of Statistical Correlations Possible? Some Theoretical Reflections Starting from the Analysis of a Conversation Corpus

There is a gap between what can be obtained by statistical analysis and the needs of linguistic theory, especially in reference to pragmatics and discourse analysis. As a matter of fact, the statistical study of correlations always points towards a causal interpretation of them. This interpretation can be explicitly aimed for, as done in computer science in recent causal analysis. One can also be more cautious when talking about causes behind correlations, unless within strictly defined experimental conditions, as traditional statisticians prefer. Still, in both approaches, the only actual viable way of interpreting data is oriented in a potential causal direction – and this also the case for data which are collected to understand human action. Therefore, human action ends up being explained only as a residual component of interacting effects, all potentially taken into account by causal models. Now, several trends of research (in linguistics, semiotics, philosophy, anthropology and psychology) highlighted the importance of taking agents' intentions into account when working on human actions and their meaning. Describing language is also a matter of understanding why people produce sentences and what they mean with them. But until now this intentional (or immanent) dimension of language production, properly characterizing it as a cognitive ability, seems to elude quantitative formalisation.

We believe that a bridge must be built to employ quantitative approaches to this end, within linguistic theorisation. Ideally, correlations in sentences, conversation exchanges and texts should not (only) be thought as the result of some causes behind language production, but as the realisation of some ends given form by language itself. We need to be able to give a finalistic interpretation to the correlations observed in data, in a way similar to what is already done in causal analysis. For example, whenever a correlation between a question and an answer is observed, we would need a sort of test capable of showing whether or not the answer was the intended effect of the question (as in similar situations we say that a certain effect was due to a certain cause).

This presentation will start from the results of the analysis of an online linguistic exchanges corpus, modelled using a bottom-up technique called process mining. We have already identified some patterns in the expression of intentions. However, apart from the interest of these patterns in themselves for the study of conversation, can they be seen as the backbone for identifying traces of intentional activity?

ALIZÉE LOMBARD, LUCIE BARQUE, DORIANE GRAS, RICHARD HUYGHE (Université de Fribourg)

Sens nouveaux et degrés de régularité : une approche quantitative du sentiment néologique

Le sentiment néologique, i.e. l'habileté à déceler la nouveauté lexicale, est encore très peu étudié expérimentalement (Gardin et al. 1974, Sablayrolles 2003, Frisson and Pickering 2007, Rodd et al. 2012). Il a été soutenu qu'il dépend d'au moins deux variables linguistique : la (non-)nouveauté de la forme et la (non-)régularité du procédé qui a produit le néologisme.

Notre étude est focalisée sur cette dernière variable et sur les néosémies, i.e. les mots dont seul le sens est nouveau (ex. *tsunami* dans *un tsunami de touristes*). Ces derniers relèvent de différents procédés sémantiques (métaphore et métonymie), obéissant eux-mêmes à des règles d'instanciation plus ou moins régulières (Apresjan 1974, Barque 2008), au sens où elles sont plus ou moins productives dans le lexique. Le principe de contiguïté référentielle qui permet de désigner une pièce par un nom de meuble (*bar, bureau*) est moins régulier que celui qui permet de désigner une portion standard de substance alimentaire par la substance elle-même (*yaourt, bière, chocolat, glace, café, sucre, sirop*, etc.). Nous distinguons ainsi deux degrés de régularité des polysémies et faisons l'hypothèses qu'ils déterminent les jugements métalinguistiques des locuteurs. Plus un sens nouveau est régulier, moins il a de chance de provoquer un sentiment néologique.

Pour vérifier cette hypothèse, nous présentons 84 phrases simples contenant un mot cible à deux groupes de locuteurs natifs du français, des étudiants entre 18 et 30 ans, invités à indiquer si la phrase contient « un mot employé dans un sens nouveau » et à prononcer le mot en question. Selon le groupe, les mots cibles sont utilisés alternativement avec leur sens original ou avec un sens nouveau créé pour l'expérience, par métaphore ou par métonymie. Nous observons le taux de mots repérés par chaque participant, le temps de réponse ainsi que le temps de fixation et l'existence ou non de retours sur le mot cible durant la lecture, mesurés par suivi oculaire. Nous postulons qu'un sentiment néologique fort se manifeste par un taux de repérage élevé, un temps de réponse rapide, un temps de fixation long et des retours en arrières. Nous testons la corrélation entre ces mesures et les propriétés des néologismes, avec comme hypothèses que

(i) les néosémies seront plus repérées que les sens originaux,
(ii) les néosémies de faible régularité provoqueront des temps de réponse plus courts que celles de forte régularité,
(iii) les temps de fixation seront plus long pour les néosémies que pour les sens originaux,
(iv) les retours en arrière seront plus importants pour les néosémies que pour les sens originaux,
(v) les effets observés en (i), (iii) et (iv) seront plus importants pour les néosémies de faible régularité que celles de forte régularité.

Avec une approche expérimentale quantitative à l'interface de la linguistique et de la psychologie cognitive, notre étude vise à contribuer aux théories linguistiques à propos du sentiment néologique. Elle pourrait en effet permettre une meilleure compréhension de la manière dont nous appréhendons et construisons les sens nouveaux, et plus généralement des processus cognitifs de traitement sémantique.

Références

Apresjan, J. (1974). Regular Polysemy. In : Linguistics (42): 5-32.

Barque, L. (2008). Description et formalisation de la polysémie régulière du français. Thèse en vue del'obtention du diplôme de docteur de l'Université Paris 7.

Frisson, S. and Pickering, M. (2007). The processing of familiar and novel senses of a word: Why reading Dickens is easy but reading Needham can be hard. Language and Cognitive Processes, 2(4), 595–613.

Gardin, B., Lefèvre, G., Marcellesi, C., and Mortureux, M.-F. (1974). A propos du « sentiment néologique ». In: Langages (36): 45-52.

Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., and Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. Memory and Cognition, 40(7), 1095-1108.

Sablayrolles, J.-F. (2003). Le Sentiment néologique. In: L'Innovation lexicale. Dir. J.-F. Sablayrolles. Paris, Champion: 279–295.

TRISTAN PURVIS (American University of Afghanistan)

Variation in relative clause constructions in Dagbani

Dagbani employs head-internal relative constructions (HIRC) for the formation of object relatives, as found in other Gur languages (see, e.g., Hiraiwa 2009), whereas subject relatives follow a more typical head-external construction employing relative pronouns that are derived from the class of strong pronouns. (These are exemplified further below in (1) and (2).) Wilson (1963, 1972) has provided a relatively thorough sketch of these constructions and their mutually exclusive distribution, and the limited accounts of Dagbani written since then typically refer to Wilson's analyses. However, there are a number of points left unresolved or unaddressed.

With object relatives, for example, the relativized noun may either be left in its canonical sentence position or positioned at the front of the relative clause.

(1) a. bia so o ni yɛn dɔɣi maa gba nyɛ -la . . . [*ex-situ head noun*]
child INDEF 3S SUB aboutto bear DEF also COP FOC
'The child whom she is going to bear also is …' (source: tech. medical pamphlet)

b. . . . bɛ ni yɛn dɔɣi bi' shɛba maa gba nya . . . [*in-situ head noun*]
3P SUB aboutto bear children INDEF DEF also get
'… the children to whom they are about to give birth also get …' (source: writ. conv./med.)

Wilson (1963) speculates that the choice between these variant forms may be largely motivated by the position of the noun in relation to the verb of the main clause, while Olawsky (1999) comments anecdotally that the fronted variety is more common in his data.

With subject relatives, Wilson notes that the so-called indefinite pronouns—which are said to be a required component for object relatives (as seen in (1))—are optional, as exemplified in (2).

(2) a. . . . bi' so ŋun be o puli ni maa [*subj. rel. w/ indef.*]
child INDEF REL belocated 3S stomach LOC DEF
'… the child who is in her stomach' (source: MEDICAL-SPOKEN RECORDING)
b. . . . bihi ban ka kom viɛnyɛliŋga bɛ niŋgbuna puuni [*subj. rel. w/o indef.*]
children REL nothave water well 3P bodies inside
'… children who don't have enough water in their bodies' (source: medical pamphlet)

In this paper I employ a combination of quantitative analysis and qualitative examination of examples to explore the degree to which these variant forms may be predicted by a number of contextual factors including sentence position, grammatical role, focus status, animacy, medium of communication, and genre, based on patterns found in a corpus of 163 texts excerpts (as exemplified above) totaling roughly 140,000 words. While primarily approached as a descriptive work, documenting additional aspects of relative clause constructions that have not been fully covered in previous grammatical accounts of the Dagbani language, the results have clear implications for theoretical considerations, and this study thus illustrates how corpus methods can contribute to theoretical linguistics in general, including research on language variation more so in this particular case.

References

Hiraiwa, K. 2009. A note on the typology of head-internal relativization. In P. K. Austin, O. Bond, M. Charette, D. Nathan & P. Sells (eds) *Proceedings of Conference on Language Documentation and Linguistic Theory 2*, pp. 123-131. London: SOAS.

Olawsky, K. J. 1999. *Aspects of Dagbani Grammar, with special emphasis on phonology and morphology*. Muenchen: LINCOM Europa.

Wilson, W. A. A. 1963. Relative constructions in Dagbani. *Journal of African Languages, 2*(2), 139-144.