Workshop "Corpus-driven studies of heterogeneous and multilingual corpora"

Paris-Villejuif, 26 September 2016

9h30 – Welcome

Introduction
Evangelia Adamou (CNRS, Lacito) & Isabelle Léglise (CNRS, SeDyL)

10h Invited speaker
Felicity Meakins (Univ. of Quennsland)
Breadth and depth: Quantifying complexity across populations in situations of language contact

11h Sophie Alby (Univ. Guyane, SeDyL)
Interactional annotation of plurilingual and heterogenous corpora

11h45 Gudrun Ledegen (Rennes 2, Prefics)
Code-switching in Reunionese SMS: a continua from minimal to intense practices, in a 'floating' milieu

12h30 Buffet lunch

14h Case studies

14h Djegdjiga MOHDEB-AMAZOUZ (LPP, Paris Sorbonne Nouvelle), Martine ADDA-DECKER (CNRS, LPP), Lori LAMEL (CNRS, LIMSI),
Arabic-French code-switching across Maghreb Arabic dialects: a quantitative analysis

14h45 Stefano Manfredi (CNRS, SeDyL)
Migration and lexifier-creole language contact: The case of Juba Arabic

15h30 Chrystelle Talla (Univ. Yaoundé)
Deciphering a hybrib sociolect result of language contact in multilingual Cameroon: Camfranglais

Workshop "Corpus-driven studies of heterogeneous and multilingual corpora"

Where: CNRS campus at Villejuif, France

When: September 26, 2016

Convenors: Evangelia Adamou and Isabelle Léglise

In the past ten years, there has been a growing number of spoken corpora annotated for language contact including for lesser-described languages (Adamou 2016) and for multilingual settings illustrating all sorts of bilingual talk, codeswitching, heterogeneous language practices or polylanguaging (Léglise & Alby 2013, 2016).

This workshop will focus on the quantitative analysis of free-speech corpora in various settings and what they bring for the description of mixed languages, bilingual talks and mixed or heterogeneous language practices. Indeed, a quantitative approach of spoken corpora has demonstrated the creation of special outcomes of codeswitching, such as "mixed languages" (McConvell & Meakins 2005), and so-called "unevenly mixed languages" (Adamou & Granqvist 2015). Moreover, the quantitative approaches of mixed and heterogeneous corpora of typologically diverse languages can re-evaluate the relevance of the very influential "matrix" and "numerically-dominant language" concepts (Myers-Scotton 1993, Muysken 2000).

More specifically, this workshop aims at bringing together researchers addressing the following questions:

- The degrees of language mixing based on a quantitative analysis of spoken corpora.
- The degrees of heterogeneity within corpora.
- Typologies of bilingual talk.
- Language mixing typologies.

Funding:

This workshop is part of the project "Multifactorial analysis of language changes" http://axe3.labex-efl.org/?q=fr/LC1f (PI Isabelle Léglise) funded by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (ANR-10-LABX-0083) *Empirical Foundations of Linguistics*.

Felicity Meakins

Breadth and depth: Quantifying complexity across populations in situations of language contact

One of the oft claimed results of language contact is the reduction of complexity, in particular what Anderson (2015: 20) refers to as "complexity of exponence". For example, syncretism, allomorphic simplification, the un-transferability of morphemes and increased paradigmatic regularity are all observed outcomes of contact-induced change (e.g. Gardani, 2008; Janse & Tol, 2003; McWhorter, 1998; Myers-Scotton, 2002). These processes reduce the exponence of morphological features, e.g. case, TAM, gender, number; and the complexity of relationships *between* cells within paradigms expressing these features. Such changes can be quantified as a measure of 'entropy' i.e. the degree of predictability between cells in a paradigm (Ackerman, Blevins, & Malouf, 2009).

One area of complexity, which Anderson (2015: 22) notes as having received less attention in the morphological literature, is variation *within* the cells of a paradigm, e.g. 'dived' and 'dove' as different word forms of the past tense form of {DIVE} in English. Thornton (2011) terms this type of complexity, morphological "overabundance", which refers to multiple forms being realized in the same cell of a paradigm. She observes that variation between forms is usually motivated by different phonological and syntactic-semantic conditions.

This paper discusses "overabundance" as one area of morphology which proliferates in situations of language contact and, in fact, represents complexification. It argues that variation in the expression of functional categories requires speakers to make calculated choices about morphemes based on other features in the clause. This paper shows how overabundance can be measured across a speaker population using mixed modelling for single morphological variants (deep measure), and the Wright-Fisher population genetics model for multiple variants (broad measure). Two case studies are presented using corpus of Gurindji Kriol, which is a mixed language spoken in northern Australia, totalling 80hrs (57,179 clauses) from 70 speakers (Meakins, 2011).

Case study 1: Deep measures of morphological overabundance (with Sasha Wilmoth)

Overabundance in Gurindji Kriol manifests itself as optional subject marking. It involves variation within a cell in a paradigm, i.e. the (non-)use of a case suffix where the grammatical role of the nominal is unaffected by non-use. Optional subject marking developed when the Gurindji ergative marker was retained in the process of the formation of the mixed language, but also came to mark intransitive subjects (Meakins, 2015). In Gurindji, the ergative marker was grammatically obligatory, whereas in Gurindji Kriol, the nominative (<ergative) became optional. Therefore, morphological overabundance developed in the nominative cell where an alternation now exists between the forms *-ngku/-tu* and a zero morph. The variation is driven by a number of semantic, syntactic, and information structure features including animacy of the agent and word order.

Thus, overabundance in Gurindji Kriol is an example of a contact-induced change which involves the complexification of an inflectional paradigm rather than its simplification. This type of complexity can be measured using generalized linear mixed models which probabilistically measure the use vs non-use of a feature (dependent variable) against semantic, grammatical and information structure features in a clause (independent variables) and their interactions, within a cluster of idiolects (random variable) (Baayen, 2008; Marschner, 2011; Pinheiro & Bates, 2000). The measure of complexity is the number of semantic, grammatical and information structure features (variables) required for a model to fit and reach a reportable R^2 value.

Case study 2: Broad measures of morphological overabundance (with Xia Hua and Lindell Bromham)

The previous case study measures overabundance for one variant morpheme in Gurindji Kriol. In order to describe variation across multiple variables, we adapted the Wright-Fisher model (Fisher, 1956; Wright, 1931) which is a classic model in population genetics that describes changes in allele (read: linguistic variant) frequencies over generations. The analogies between changes in allele frequencies and changes in linguistic variants stem from the fact that the transmission of an allele from parent to offspring and a person reproducing a linguistic variant that he/she has heard before are both replication processes (Blythe, 2012). Several theoretical studies have examined the mathematical link between the Wright-Fisher model and models of language evolution (e.g. Baxter, Blythe, Croft, & McKane, 2006; Bentley, Ormerod, & Batty, 2011; Reali & Griffiths, 2010), however this is the first time this model has been applied to language data. We explore population-level variation with respect to changes in morphological complexity across the entire language system, but also language dominance (Kriol or Gurindji), and the fixation of language features (i.e. reduction of variation, and therefore simplication) across 250 variables in 70 people. All variables represent a paradigmatic choice (morphological, but also lexical in this study) between Gurindji or Kriol forms.

Sophie Alby

Interactional annotation of plurilingual and heterogenous corpora

This communication aim is to explain the methodological choices that have been made to annotate contact phenomena in the Clapoty project and more specifically the interactionally “remarkable” phenomena (PRINT). The challenge for interactional phenomena was to annotate what is remarkable at the level of the speech turns or conversational sequences in terms of language choice. To do so, we made the choice of a structural approach of interactions based on Auer (1995) model in order to describe the function of languages choices and to highlight the underlying “grammar” (Matthey and De Pietro, 1997) of these plurilingual interactions, to show how languages are intentionally and context-bound used by humans (Jørgensen et al. 2011). The level of analysis is a pragmatic one as it aims to explain how speakers give meaning to their speech and structure it.

Gudrun Ledegen

Code-switching in Reunionese SMS : a continua from minimal to intense practices, in a 'floating' milieu

La comparaison des pratiques alternantes entre le terrain affinitaire à La Réunion (Ledegen 2015) et le terrain francophone en Suisse (Morel, 2016) fait apparaître des contrastes mais aussi des proximités éclairantes sur l'alternance codique dans le cadre des SMS :

- d'une part, la situation franco-créolophone se différencie nettement de la situation suisse, par l'omniprésence de la zone 'flottante' (Ledegen, 2007, 2012), zone d'interprétation multiple pouvant être associée autant au(x) français (de La Réunion, de métropole) qu'au créole réunionnais ; par ailleurs, la pratique alternante est très présente, s'inscrivant dans la logique des *mixed languages* (Auer 1999) pour les uns et du contrastant *code-switching* (Auer 1999) pour les autres. L'analyse fréquentielle de ces différentes pratiques éclairera ce contraste.

- d'autre part, des pratiques minimales d'alternance, sollicitant un panel plus large de langues, se révèlent étonnamment proches, et permettent d'inscrire au sein dudit panel, pour la situation réunionnaise, les langues affinitaires que sont le français et le créole. Cette pratique minimale, qui se réalise dans des structures formulaïques qui constituent un acte de langage, se retrouve dans de multiples situations d'apprentissage ('le premier étage de la fusée') et vient interroger la classification *alternation/insertion* (termes français proposés par S. Pekarek-Doehler (2011) pour *alternational/insertional code-switching* de Auer (1996)). Cet « élément isolé (lexème ou expression figée) qui a valeur d'énoncé/acte pragmatique (salutation, adieu, …) » (Pekarek Doehler, 2011 : 57) constitue en effet autant une *alternation* qu'une *insertion* (Muysken 2000). Pourrait-il s'agir d'une 3° catégorie à ajouter à la classification du code-switching ?

Djegdjiga MOHDEB-AMAZOUZ, Martine ADDA-DECKER, Lori LAMEL

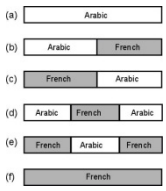
Arabic-French code-switching across Maghreb Arabic dialects : a quantitative analysis

Code-switching (CS), i.e. the dynamic switching from one language to another within a given oral or written speech interaction, is a phenomenon resulting from language contact. Arabic-French code-switching is relatively frequent in Maghreb countries, although more typical of Algerian Arabic [1, 3]. French, considered by Algerian speakers a prestigious language and the language of highly educated people [1], mixes with the Algerian Arabic dialect, mother tongue and everyday language of Algerian speakers. In this context, the two languages may come into concurrent [4] or complementary use which leads to a particular practice of CS. We aim at quantifying the frequency of Arabic-French CS across three Arabic dialects: Algerian, Moroccan and Tunisian. We propose to analyse the number of segments including CS. Our goal is to rst quantify the switch segments and then classify these

segments in terms of CS quantity and order. Using this switch order and frequency of both languages, we de ne the matrix language and embedded language[2] of the segments.

Our corpus consists of 53 hours of mostly entertainment TV shows and talk shows from stations in Algerian dialects (14 hours), Moroccan (15 hours), Tunisian (24 hours). We are particularly interested in the presence of CS within breath groups. We manually segmented all the data into breath groups, hence-forth called segments. These segments were then labelled using their respective language or CS labels specifying language order and number of changes. As shown in the table Figure 1 whereas most of segments are Arabic, there are 2850 purely French segments. Most segments with more than one language start with Arabic and switch to French. A total of 4248 segments including French in the 53 hours data set results in an average CS density of 80 segments per hour.

Table 1 shows details across our three dialects, with the total number of speech segments including French and the full show durations per dialect. A CS density is computed as the ratio between French colored segments and total duration. Whereas the Moroccan and Tunisian dialects include about 35 CS segments per hour, the Algerian CS density raises to 200 segments per hour, highlighting the importance of Arabic-French CS in the Algerian Arabic dialect.



Segment labels	nb.segments	dur. (hh:mm:ss)
Arabic	65844	50:18
Arabic-French	877	00:50
French-Arabic	359	00:17
Arabic-French-Arabic	157	00:11
French-Arabic-French	5	00:00:17
French	2850	1:18

Figure 1: The manually isolated speech segments (breath groups) were anno-tated with six major labels: (a) fully Arabic, (b) CS Arabic-French, (c) CS French-Arabic,(d) CS Arabic-French-Arabic, (e) CS French-Arabic-French, (f) fully French. The last 2 columns give the total number of segments (and dura-tion) per segment type (all dialects pooled).

Dialect	nb. CS segments	total show dur.	nb. CS/hour
ALG	2801	14h	200.1

TUN	938	24h	39.1
MAR	509	15 h	33.9

Table 1: Details across three dialects: Algerian, Tunisian, Moroccan: total number of CS segments involving French, full show duration and resulting CS density (number of CS segments per hour).

Stefano Manfredi

Migration and lexifier-creole language contact: The case of Juba Arabic

Juba Arabic is an Arabic-based pidgin-creole mainly spoken in Juba, the capital of South Sudan. The sociolinguistic situation of Juba Arabic involves pidginization and creolization going on simultaneously, since there is a constant influx of both native and non-native speakers of the language (Manfredi forth.). In this overall situation, the pidginized varieties undergo grammatical stabilization because of the gradual functional expansion of Juba Arabic, whereas the creolized varieties are more or less affected by phonological and morphosyntactic restructuring because of the longstanding exposure of the speakers to the lexifier language (i.e. Sudanese Arabic). Depending on in-group sociolinguistic factors which determinate the amount of the exposure to Sudanese Arabic, Juba Arabic is thus affected by a gradual process of levelling towards its lexifier language (Versteegh 1993). Based on an ongoing contrastive study of two corpora recorded in Juba and in Khartoum (Manfredi 2014), this presentation aims at revealing how different migration trajectories of Juba Arabic speakers induce different outputs in the creole-lexifier language contact. The analysis will focus on a number of morphosyntactic features such as the morphological encoding of subject, TAM marking and nominal determination. Above and beyond, the study seeks to contribute to the enduring theoretical discussion concerning the definition of the synchronic process of decreolization (Aceto 1999, Goury & Léglise 2005, Siegel 2010).

References

Aceto, M. 1999. Looking beyond decreolization as an explanatory model of language change in creole-speaking communities. Journal of Pidgin and Creole Languages 14 (1): 93-119.

Ackerman, F., Blevins, J., & Malouf, R. (2009). Parts and wholes: Patterns of relateness in complex morphological systems and why they matter. In J. Blevins & J. Blevins (Eds.), *Analogy in Grammar: Form and Acquisition* (pp. 54-82). Oxford: Oxford University Press.

Adamou, E. & Granqvist, K. (2015). Unevenly mixed Romani languages. *International Journal of Bilingualism* 19(5): 525-547.

Adamou, E. (2016). *A corpus-driven approach to language contact. Endangered languages in a comparative perspective*. Boston & Berlin: de Gruyter Mouton.

Anderson, S. (2015). Understanding and Measuring Morphological Complexity. In B. Matthew, B. Dunstan, & C. Greville (Eds.), *Understanding and Measuring Morphological Complexity* (pp. 11-26). Oxford: Oxford University Press.

Auer P., 1995, "The pragmatics of code-switching: a sequential approach", in Milroy, L. & Muysken, P. (Eds), *One Speaker, Two Languages*. *Cross-Disciplinary Perspectives on Code-Switching*, Cambridge, Cambridge University Press, 115-135.

Auer P., 1999, "From codeswitching via language mixing to fused lects : toward a dynamic typology of bilingual speech", *International journal of bilingualism* III-4, 309-322.

Baayen, H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baxter, G. J., Blythe, R. A., Croft, W., & McKane, A. J. (2006). Utterance selection model of language change. *Physical Review E, 73*, 046118.

Bentley, R. A., Ormerod, P., & Batty, M. (2011). Evolving social influence in large populations. *Behavioral Ecology Sociobiology, 65*(537−546).

Blythe, R. A. (2012). Neutral evolution: a null model for language dynamics. *Advances in Complex Systems, 15*, 1150015.

Fisher, R. A. (1956). *The genetical theory of natural selection*. New York: Wiley.

Gardani, F. (2008). *Borrowing of inflectional morphemes in language contact*. Frankfurt: Peter Lang.

Goury, L. & Léglise I., 2005. Contact de Créoles, Créoles en Contact. Paris: L'Harmattan.

Janse, M., & Tol, S. (Eds.). (2003). *Language Death and Language Maintenance: Theoretical, Practical and Descriptive Approaches*. Amsterdam: John Benjamins.

Laroussi, F. (1997). Plurilinguisme et identités au Maghreb, vol. 233. Publication Univ Rouen.

Ledegen, G. & Richard, M., (2007), « « *jv me prendre un bois monumental the wood of the century g di* ». Langues en contact dans quatre corpus oraux et écrits « ordinaires » à la Réunion », *Glottopol*, n° 10, 'Regards sur l'internet, dans ses dimensions langagières. Penser les continuités et discontinuités', 86-100.

Ledegen, G., (2012). « Prédicats "flottants" entre le créole acrolectal et le français à La Réunion : exploration d'une zone ambigüe », dans Chamoreau, C. & Goury, L. (Eds), *Systèmes prédicatifs des langues en contact,* CNRS Editions, Coll. « Sciences du langage », 251-270.

Ledegen, G., (2015), *La dimension « flottante » dans le contact de langues : analyses syntaxique & sociolinguistique d'un grand corpus de pratiques ordinaires orales et écrites à La Réunion*, Habilitation à Diriger des Recherches, Université de Rennes 2.

Léglise, I. & Alby S. (2013). Les corpus plurilingues, entre linguistique de corpus et linguistique de contact. *Faits de Langues* 41: 95-122.

Léglise, I. & Alby S. (2016). Plurilingual corpora and polylanguaging, when corpus linguistics meets contact linguistics. *Sociolinguistic studies* 10(3).

Manfredi, S. 2014. "Juba Arabic corpus". In A. Mettouchi, M. Vanhove, D. Caubet (ed.), ANR, CorpAfroAs: a corpus for spoken Afro-Asiatic Languages, http://corpafroas.tge-adonis.fr/

Manfredi, S. forthcoming. Arabic Juba: un pidgin-créole du Soudan du Sud. Leuven-la-Neuve: Peeters.

Marschner, I. (2011). glm2: Fitting generalized linear models with convergence problems. *The R Journal, 3*(2), 12-15.

McConvell, P. & Meakins F. (2005). Gurindji Kriol: A mixed language emerges from code-switching. *Australian Journal of Linguistics* 25(1). 9–30

McWhorter, J. (1998). Identifying the creole prototype: Vindicating a typological claim. *Language, 74*, 788-818.

Meakins, F. (2011). *Case marking in contact: The development and function of case morphology in Gurindji Kriol*. Amsterdam: John Benjamins.

Meakins, F. (2015). From absolutely optional to only nominally ergative: The life cycle of the Gurindji Kriol ergative suffix. In F. Gardani, P. Arkadiev, & N. Amiridze (Eds.), *Borrowed Morphology* (pp. 189-218). Berlin: Mouton de Gruyter.

Morel, E., (2016) *Le bricolage plurilingue dans la communication par texto. Interprétations d'une pratique entre affiliation locale et aspiration globale*, Thèse de doctorat en Sciences du langage, dirigée par S. Pekarek Doehler et B. Siebenhaar, Univ de Neuchâtel.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-mixing.* Cambridge University Press: Cambridge.

Myers-Scotton, C. (1993). *Duelling Languages: Grammatical Structure in Code-switching*. Oxford: Clarendon press.

Myers-Scotton, C. (2001). The matrix language frame model: Development and responses. Trends in Linguistics Studies and Monographs 126, 23{58.

Myers-Scotton, C. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford ; New York: Oxford University Press.

Pekarek Doehler S., '2011), « 'Hallo! Voulez vous luncher avec moi hut?' Le "code switching dans la communication par SMS », *Linguistik Online* 48 4/2011. http://www.linguistikonline.de/48_11/index.htm

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Reali, F., & Griffiths, T. L. (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of Royal Society B, 277*, 429 – 436.

Siegel, J., 2010. "Decreolisation: a critical review". In: J. C. Clements, M. E. Solon, J. F. Siegel, and B. D. Steiner (eds.), IUWPL9. Bloomington: IULC Publications. 83–98.

Thornton, A. M. (2011). Overabundance (Multiple forms realizing the same cell): A non-canonical phenomenon in Italian verb morphology. In M. Maiden, J. C. Smith, M. Goldbach, & M.-O. Hinzelin (Eds.), *Morphological Autonomy: Perspectives from Romance Inflectional Morphology* (pp. 359-382). Oxford: Oxford University Press.

Versteegh, K. (1993). "Leveling in the Sudan: from Arabic creole to Arabic dialect". International Journal of the Sociology of Language 99: 65-97.

Wright, G. (1931). Evolution in Mendelian populations. *Genetics, 16*, 97−159.

Zaboot, T. (2010). La pratique langagière de locuteur(s) bilingue (s)'. Synergies Algerie 9, 201-10.

Ziamari, K. (2009) Le contact entre l'arabe marocain et le francais au Maroc: spécificités linguistique et sociolinguistique. Synergies Tunisie 1, 173-186.