Journée d'Etudes AFLiCo JET : "Corpora and Representativeness"

Les journées d'étude de l'AFLiCo (Association Française de Linguistique Cognitive) sont des moments de discussion scientifique de qualité pour la recherche en linguistique cognitive et, plus généralement, les approches linguistiques fondées sur l'usage. Cette année, le thème retenu est : "Corpus et Représentativité" et elle est organisée en partenariat avec un membre de Paris8/MoDyCo/IUF et le CREA EA 370

Date(s)
du 3 mai 2018 au 4 mai 2018

Lieu(x)
Bâtiment W (Max Weber)
Salle des Conférences

Conférenciers invités:

- **Dawn Knight**, School of English, Communication and Philosophy, Cardiff University: “Representativeness in CorCenCC: corpus design in minoritised languages”
- **Thomas Egan**, Inland Norway University of Applied Sciences: “Some perils and pitfalls of non-representativeness”

Site de la JET:


Comité d'organisation:

- Guillaume Desagulier (Paris8/MoDyCo/IUF),
- Sophie Raineri (EA 370 CREA, Nanterre)
- et Camille Debras (EA 370 CREA, Nanterre)

The French Cognitive Linguistics Association (AFLiCo) invites you to submit paper proposals for its next workshop, AFLiCo JET 2018. AFLiCo workshops provide a forum for high-quality research in cognitive linguistics and, more generally, usage-based approaches to language. The topic of this year’s workshop is **corpora and representativeness**.

With the advent of corpus linguistics, the use of corpora has become central in linguistics. One underlying
assumption is that the corpus is representative of the linguistic phenomenon under scrutiny. Of course, corpus representativeness itself is a methodological construct (Leech 2006, Habert 2010): language corpora are tools constructed by linguists, and their structural limitations constrain and condition the validity of linguistic findings.

Here is an open-ended list of issues that we wish to address in the workshop:

What does it mean for a corpus to represent language use, and what are the relevant criteria?
To what extent do representativeness rely on intuition, since it cannot be fully gauged empirically? Because a corpus cannot be representative of all features of language use how can we address bias in sampling?
Does representativeness necessarily entail balance? Can the design of a corpus be totally free from any form of theorization?

Solutions to these complex issues may reflect in the development and use of different types of corpora.

The representativeness of written corpora may rely on a variety of features. According to Biber (1993: 244), “[r]epresentativeness refers to the extent to which a sample includes the full range of variability in a population.” Variability can be defined as the interaction between situational (e.g. format, setting, author, addressee, purposes, topics) and linguistic, distributional parameters (e.g. frequencies of word classes). Sampling can be based on extralinguistic (sociological, demographic) criteria (Crowdy 1993). Balance, i.e. a proportion of sampled elements that reflects their frequency in the targeted language, is claimed to characterize some corpora (e.g. the Brown Corpus (Francis & Kucera 1979) and the Lancaster-Oslo-Bergen corpus (Johansson et al. 1978)), though it is not a prerequisite.

Although increasingly larger corpora, including monitor corpora, can be compiled from the Web (Baroni et al. 2009), large size is not necessarily a priority. “Big is beautiful” in the realm of corpora is, perhaps, a “delusion” (Svartvik 1992: 10). Large corpora are often presented as an ideal but, in practice, “small” corpora can go a long way in such domains as English language teaching (Ghadessy, Henry, and Roseberry, 2001), the study of metaphors (Cameron and Deignan 2003), dialectology (Hollmann and Siewierska, 2007; Boas and Schuchard, 2012), etc. Parallel corpora, i.e. collections of original texts and their translations in one or more languages, are particularly useful in areas of research such as contrastive linguistics, translation studies and computational linguistics (Kenning 2010), but their alleged lack of representativeness has called for inventive ways of using them (Nádvorníková 2017).

In the area of spoken corpora, collecting data that represents the variability of the multiple dimensions of speech (phonology and phonetics, prosody, gesture) remains a challenge today. Collecting, transcribing, annotating and analysing data, is a slow, sometimes complicated, task. Although phonological and prosodic annotations can be partially systematized (Bertrand et al. 2008), technological advances are yet to be made in the automatic recognition of speech and gesture in interactional contexts. Automatic motion capture technologies for gesture research are promising (Priesters & Mittelberg 2013, Guez et al. 2013), but little advanced. As part of initiatives such as the TGIR Huma-Num Multi-Com – CORLI Consortium, multimodality researchers collaborate to develop collective harmonised practices of collection, transcription and archiving of spoken corpora.

References


CALL FOR PAPERS

AFLiCo JET 2018 invites linguists, including junior researchers, to submit proposal that address the following themes (this is an open-ended list):

Bias in corpora
Material issues in corpus building
Theoretical issues in corpus building
The use of different types of corpora in a complementary fashion in linguistic analysis
Balance, size, distribution in corpora representativeness
Spoken/multimodal vs written/textual corpora
Guidelines for corpus building

Anonymous abstracts for 20-minute presentations (+ 8 minutes for questions) should include a title and a short bibliography. They should not exceed 500 words (exclusive of references, tables, and figures). They can be in English or in French.

Abstracts should clearly state the following:

research question(s)
approach(es)
subfield (e.g. semantics, pragmatics, gesture studies, corpus linguistics, NLP, etc.)
data
expected or confirmed results.

Include three to five keywords specifying the (sub)field, the topic, and the approach.

Contact:
Guillaume Desagulier (Paris8/MoDyCo/IUF), Sophie Raineri (CREA, Nanterre) et Camille Debras (CREA, Nanterre): cdebras@parisnanterre.fr