Corpus Linguistics: Epistemological and Methodological Issues in Language Studies

Joint Master in European Languages, Cultures and Societies in Contact

(Research Methodologies in European Modern Languages and Literatures)

5th March 2026

Session 13

- ***1. What is Corpus Linguistics ?***
- - another way of doing linguistics, more than a methodology
- - reflection on how linguists work
- - relationship between type of data and how theory is built

- Auroux (1998), Dalbera (2002), Willems (1998)

	Linguistic theories / Types of data	Approach / Method
Traditional Grammar	Introspection language as a system	Hypothetical-deductive approach / method
Generative Grammar	Introspection + ungrammatical examples + minimal commutations – verify the possibilities of the schemas	Hypothetical-deductive approach / method
Field Linguistics (Corpus Linguistics)	Linguistic investigation - Spoken data (observation) - Written data - Language survey description of speakers' behavior in very diverse socio-linguistic contexts	Empirical-inductive approach / method

- **Hopper (1987) - “emergent grammar”**

- ‘I believe the same is true of **grammar**, which like speech itself must be viewed as a **real-time, social phenomenon, and therefore is temporal**; **its structure is always deferred**, always in a process but never arriving, and therefore **emergent**; and since I can only choose a tiny fraction of data to describe, any decision I make about limiting my field of inquiry (for example in regard to the selection of texts, or the privileging of the usage of a particular ethnic, class, age, or gender group) is very likely to be **a political decision**, to be against someone else's interests, and therefore disputed. The notion of **Emergent Grammar** is meant to suggest that ***structure, or regularity, comes out of discourse and is shaped by discourse as much as it shapes discourse in an on-going process***. Grammar is hence not to be understood as a pre-requisite for discourse, a prior possession attributable in identical form to both speaker and hearer. Its forms are not fixed templates but are negotiable in face-to-face interaction in ways that reflect the individual speakers' past experience of these forms, and their assessment of the present context, including especially their interlocutors, whose experiences and assessments may be quite different. Moreover, the term Emergent Grammar points to a grammar which is not abstractly formulated and abstractly represented, but always anchored in the specific concrete form of an utterance.'.

- **Methods of data collection**

- A) language survey (questionnaries, interviews)
- B) participant observation
- C) recording groups in speech interaction (ecological data - dinner, consultation, market, hotel, shop, phone, so on)

2. Corpus

- collection of texts, interactions, representative of a particulary language or a variety of language
- 2.1. corpus – construct

- **2.2. Written corpora**
- - texts
-
- Sketch engine
- Frantext
-
- **Spoken corpora**
- - recording interactions + transcribing
- - ***time alignment*** (COLT corpus - London teenage speech ; International Corpus of English http://ice-corpora.net/ice/ - British component)
-
- Spoken Italian – Interviews about Language and Nation, University of Oslo, http://tekstlab.uio.no/silana/corpus.html
-
- **- multimode -** British National Corpus
-
- **For the French-speaking world**
-
- - Corpus International Ecologique de la Langue Française http://ciel-f.org/
- - CLAPI http://clapi.ish-lyon.cnrs.fr
- - Corpus de Français parlé au Québec https://applis.flsh.usherbrooke.ca/cfpq/

- ***Corpus transcription***

- « **transcription is a selective process reflecting theoretical goals and definitions** » (Ochs, 1979)
- - corpus of Acadian French from Nova Scotia

- ***- ben 'well' / bien***

- **- *pis* 'and' /vs/ *puis***
- - lexicalization

- (15) H[6]1 : quoi-ce ça veut dire ça
- (16) F[6]2 : ***ben*** / ça que ça veut dire / c'est que je voulons commencer un groupe / yu-ce-que les femmes entrepreneurs ***pis*** les femmes professionnelles de la région de Clare / pouvont se réunir / en réunion / pour partager / pour collaborer / euh : pour apprendre du STUFF / pour : / vraiment avoir un endroit à se réunir / à cause que des femmes entrepreneurs là c'est : / c'est tout à fait unique hein

- **Corpus annotation**

- ***2.3. What the corpus is used for (“corpus based /vs/ “corpus driven")***
- “corpus based /vs/ “corpus driven" (Tognini-Bonelli, 2001)

- **2.4. *Monitor corpora /vs/ balanced corpora /vs/ opportunistic corpora***

- **Monitor corpora**
- COCA - Corpus of Comporary American English
- https://www.corpusdata.org

- Sketch engine

- WordBanks online (derived from Bank of English)

- Frantext

- **Balanced corpora (sample corpora)**
- - Lancaster Oslo/Bergen corpus (so-called LOB corpus) - standard written modern British English in the early 1960

- https://varieng.helsinki.fi/CoRD/corpora/LOB/index.html

- **Opportunistic corpora**

- **3. How to work on corpora ?**
- - concordancers
- - frequency data

- *I was going to say* (COCA)
- *J'allais dire* (TenTen)
- *Era să spun* (Romanian Web)

- **3.1. What linguistic facts / features for what type of corpus ? or**
- **What type of corpus for what linguistic facts / features ? and what kind of corpus approach (corpus-based or corpus-driven) ?**

- **Petraș (2019)**

- **Je vas dire, on va dire**

- ***Acadian folktales***

- Il arête icite, c'était un stable, ils aviont, *on va dire* un hôtel puis il y avait un stable, les chevals puis tout ça (coll. Raymonde Roussel, bob. 1a, enreg. 2, p. 4, 1977)

- Tchin, bétot i'avait une p'tit bâtisse, une p'tite shack là *on va dire*, une p'tite bâtisse à côté du ch'min [...] (coll. Carmen LeBreton, bob. 7, face B, no. 44, p. 2, 1971)

- I avait... avait soif. "Ooh ! mon Djeu, si c'était pas pour te draguer, j'irais ben boire une gorgée de bière. Le rhum *je pourrais dire*" (coll. Labelle-Richard, bob. 2, enreg. 15, p. 2, 1992)

- ***on va dire = shall we say***

- **3.2. Working on comparable corpora**

- **Lansari (2020)**

- **Sketch Engine**
- *I was going to say* (COCA) (enTenTen21)
- *J'allais dire* (frTenTen21/23)
- *Era să spun* (Romanian Web, roTenTen21)
- *Yo iba a decir que* (spanish, esTenTen23)
- *Stavo per dire* (italian, itTenTen20)

- **3.3. Epistemological issues**

- ***- asteure 'now'*** (Petraș, 2021)
-
- **- manière (de)**
-
- *manière de rouge* 'kind of, like'
-
- 359) F[1]3 : ben vraiment asteure là / c'est ***manière de*** un petit jeu là chez nous / alle a des petits bébés partout partout là
-
- (54) F[2]2 : c'est ça / ça c'est ***manière de*** spécial là
-
- (64) F[2]2 : y a deux FERRY / i faut que tu prennes pour te rendre à l'hôpital ou au médecin à Digby ça fait
- (65) F[2]1 : ouais
- (66) F[2]2 : qu'il ont amené ça fait depuis / euh : Brier Island je crois qu'alle est là depuis janvier deux mille trois
- (67) F[2]1 : ouais
- (68) F[2]2 : la NURSE PRACTITIONER / BUT c'est une situation ***manière de*** différente parce qu'alle est vraiment tout seule

References

- Auroux, Sylvain (1998), « Les enjeux de la linguistique de terrain », *Langages*, 129, pp. 89-96.
- Dalbera, Jean-Philippe (2002), « Le corpus entre données, analyse et théorie », *Corpus* [En ligne] (« Corpus et recherches linguistiques »), 1, mis en ligne le 15 décembre 2003, consulté le 25 février 2026. URL : http://corpus.revues.org/10.
- Hopper, Paul (1987), « Emergent Grammar », *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, pp. 139-157.
- Lansari, Laure (2020), *A Contrastive View of Discourse Markers. Discourse Markers of Saying in English and French*, Cham, Palgrave Macmillan.
- Ochs, Elinor, 1979, « Transcription as Theory », in Elinor Ochs et Bambi Schieffelin (éd.), *Developmental Pragmatics*, New York, Academic Press, pp. 43-72.
- Petraş, Cristina (2021) « *Asteure*, archaïsme et changement linguistique : éclairages réciproques des français nord-américains et des français de France », *Linx*, 82, https://doi.org/10.4000/linx.8025.
- Petraş, Cristina (2019), « Les expressions métadiscursives dans les contes acadiens de tradition orale », *Studii de lingvistică*, 9-2/2019, pp. 201-224.
- Tognini-Bonelli, Elena (2001), *Corpus Linguistics at Work*, Amsterdam, John Benjamins.
- Willems, Dominique (1998), « Données et théories en linguistique: réflexions sur une relation tumultueuse et changeante », in Mireille Bilger *et al*. (éd.), *Analyse linguistique et approches de l'oral. Recueil d'études offert en hommage à Claire Blanche-Benveniste* (Orbis / Supplementa. Monographies publiées par le Centre International de Dialectologie Générale, Louvain, tome 10), Leuven, Paris, Peeters, pp. 79-87.