Building Diachronical Reference Corpora for the French Language

Alexei Lavrentiev

alexei.lavrentev@ens-lyon.fr

ICAR Research laboratory
CNRS, Université de Lyon

Corpora 2013
St-Petersburg, 25-27 June 2013

Outline

- History of the French diachronical text corpora or collections
- Text typology variables
- Text selection constraints
- Recent or current projects
  - CoRPTeF
  - GGHF
  - Presto
- Towards a French national corpus ?

What is a corpus?

- Various definitions exist
  - e.g. Sinclair
- Some scholars believe that the question is irrrelevant
- Wide sense: any collection of language data used for research purposes
  - One condition: sufficient metadata to know what the corpus represents and whether it is suitable for the research

Some history

- TLF (Trésor de la la Langue Française) → Frantext
  - 1960 creation of the CRTLF research center
    - later becomes InaLF, currently ATILF
    - Paul Imbs → Bernard Quemada → Robert Martin
  - 1971 – 1994 publication of the dictionary
  - 1970s creation of the text base for the TLF
    - 1000 texts → 4000 texts
    - 16th – 20th century (initial project: from the 9th cent.)
    - 90% fiction/philosophy vs. 10% academic/technical
    - no seek of representativeness or balance
  - 1980s “Discotext” CD
  - 1998 online access http://www.frantext.fr

Some history

- Dictionnaire du Moyen Français (DMF)
  - 1980s a project by Robert Martin (INaLF)
  - based on texts from 1430 to 1500
  - “extension” of Frantext

Some history

- Amsterdam corpus
  - early 1980s, dir. by Anthonij Dees
    - charters
    - literary texts (300 samples), > 3 000 000 occ.
  - used to create a linguistic Atlas of the Old French (1987)
    - re-vived in 2006 by A. Stein & P. Kunstmann

Some history

- Base de Français Médiéval
  - http://txm.bfm-corpus.org
  - Founded in 1989 by Ch. Marchello-Nizia
  - 9th – 15th centuries (exchanges with the DMF)
  - 75 texts / 3 300 000 words
    - soon 130 texts / 4 500 000 words
  - “Pragmatical” text selection policy
    - doctoral students’ work
    - exchanges
    - research projects
  - Detailed text description and typology
  - Tools for corpus building and analysis

Text typology variables

- "Textual unit"
  - Identified by author + date + intellectual content
  - Complex cases
    - *Roman de la rose:* 1st part by G. Lorris (~1227), 2nd part by J. de Meun (~1274) → 2 text units
    - *Lais* by Marie de France → one text unit with subdivisions
    - Charters, Proverbs, Songs...
  - Some choices are pragmatical rather than methodological
    - grouping charters of proverbs

Text typology variables

- Date
  - “original text” vs. manuscript
  - Precision (year... decade... century)
    - rules to establish “not before” and “not after” dates
- Region
  - author’s dialect
  - layers of scribal interventions
    - No actual text is represents purely a single dialect

Text typology variables

- Form
  - Verse / Prose / Mixture / Gloss
- Domain (médieval French)
  - Literary
  - Historical
  - Religious
  - Juridical
  - Scientific or Didactical
  - Acts of practice
    - Multiple domains possible
    - Other domains appear later (epistolary, journalism...)

Text typology variables

- Genre
  - Great number and diachronical variation
  - Depend on domain (but some are present in several domains)
  - Traditional names vs. modern classifications
- Internal text variation
  - “text planes”
    - Direct speech
    - Discourse types (narrative, descriptive, argumentative, explicative, dialogue) [Adam 1992]

Text selection constraints

- Availability
  - e.g. very few French texts before 1100
- Quality of the editions
  - “best copy” vs. “reconstruction” methods
- Quality of digitization
  - “plain text” vs. critical apparatus
  - more or less standard encoding schema
- Copyright issues
  - depending on countries critical editions of old text may or may not be subject to copyright protection

CoRPTeF

- Representative corpus of the first French texts (9th - 12th centuries)
  - http://corptef.ens-lyon.fr
  - resp. C. Guillot (ENS - Lyon)
  - funded by the ANR agency (2008-2010)
  - work on text descriptors (see online documentation)
  - further development: bilingual Old French / Medieval Latin corpus

GGHF

- Comprehensive Historical Grammar of the French Language
  - resp. S. Prevost, B. Combettes et Ch. Marchello-Nizia
    - a dozen of contributors
  - destined to replace Brunot's classical work
  - corpus used to verify and to provide examples to the previous scholarship
  - core vs. additional corpus
    - 200 000 – 250 000 words per century in the core corpus
  - texts gathered from various sources (BFM, Frantext, BVH...)
  - TXM search and analysis engine

PRESTO

- Development of the French prepositional system
  - resp. Denis Vigier (Lyon University) and P. Blumenthal (Köln University)
  - funded by ANR and DFG agencies (2013-2016)
  - substantial extension of the GGHF core corpus
  - development of tools for automated lemmatization for all periods of the history of French

Conclusion

- Towards a National corpus of the French language?
  - Frantext including texts starting from the 12th century
  - Two workshops on the feasibility held in 2012 and 2013 by the ILF research federation...
    - complex methodological issues
  - Corpus research infrastructure funding discipline specific corpora consortia created in 2011
    - 3 partly overlapping consortia: spoken / written for linguists / for humanities
  - DARIAH and CLARIN european infrastructures
  - … may be not a single reference corpus but a number of “trusted” specific corpora

Thank you!