Corpus Linguistics: history

Karën Fort
karen.fort@inist.fr

November 18, 2011

Introduction

Sources

Most of this course is largely inspired by:

- Corpus Linguistics, Edinburgh University Press [Mc Enery and Wilson, 1996]
- Sylvain Pogodalla's course on the same subject [http://www.loria.fr/ pogodall/enseignements/TAL-Nancy/notes-2008-2009.pdf],
- Cédrick Fairon's and Anne Catherine Simon's (Université de Louvain) course: Méthodologie de l'analyse de corpus en linguistique.

A history as a body of academic folklore...

"*Corpus linguists study real language, other linguists just sit at their coffee table and think of wild and impossible sentences*"

vs

"*A corpus can't describe a natural language entirely*"

A history as a body of academic folklore...

- Before 1950s: Early corpus linguistics (field linguistics)
- 1950s: Chomsky
- 1960s-now: Modern corpus linguistics

Overview of corpus-based studies 1/2

- Language acquisition: parental diaries (1876-1926) [Preyer, 1889], large sample studies (1927-1957) [Mc Carthy, 1954], longitudinal studies (1957-) [Brown, 1973]
- Spelling conventions: study of the frequency distribution of letters and letter sequences in German (stenography), 11 million words, 5,000 analysts [Käding, 1897]
- Language pedagogy: corpus for research on foreign language learning, vocabulary lists, word frequency

Overview of corpus-based studies 2/2

- Comparative linguistics: comparison of word sense in different languages [Eaton, 1940]. Corpora enabling the same kind of analyses was recreated only from 1996 [McEnery and Oakes, 1996]
- Syntax and semantics: descriptive grammar of English based on a corpus [Fries, 1952]. For French, [Georges Gougenheim and Sauvageot, 1956] describes a grammar based on grammatical choices and lexical frequency computed from 275 speakers.

Early corpus linguistics: a naïve approach?

- No representativeness
- No recording

→ methodology?

Early corpus linguistics: assumptions behind

- the sentences of a natural language are finite
- the sentences of a natural language can be collected and enumerated
- "*This was when linguists... regarded the corpus as the sole explicandum of linguistics*" [Leech, 1991]

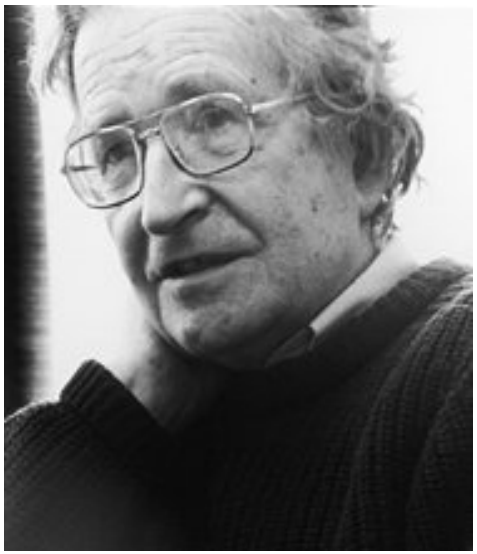
Competence vs Performance: an old debate

- rationalists vs empiricists
- empiricism: observation of naturally occurring data (corpus)
- rationalism: conscious introspective judgements.

... a VERY old debate [Rastier, 2004]

- Humboldt (1839): "in itself (language) is no product (ergon) but an activity (energeia)"
- Aristotle (350 **BC**): power vs act

Chomsky



- competence vs performance
- competence: tacit, internalized knowledge of a language
- performance: external evidence of language competence (may be influenced by other factors, though)

Chomsky

"We thus make a fundamental distinction between competence (the speaker-hearer's knowledge of his language) and performance (the actual use of language in concrete situations). [...] A record of natural speech will show numerous false starts, deviations from rules, changes of plan in mid-course, and so on. The problem for the linguist, as well as for the child learning the language, is to determine from the data of performance the underlying system of rules that have been mastered by the speaker-hearer and that he puts to use in actual performance" [Chomsky, 1965].

Example [Morrill, 2000]

Would you say that? Is it correct?

The dog that chased the cat that saw the rat that ate the cheese barked

Example [Morrill, 2000]

Would you say that? Is it correct?

The cheese that the rat that the cat that the dog chased saw ate stank

Example [Morrill, 2000]

Would you say that? Is it correct?

Yes, I **could** *say that – but I never* **would**

Chomsky's arguments against the use of corpora

- encourage to model competence rather than performance
- natural languages are **not** finite, therefore this cannot yield an appropriate description of language
- introspection must not be eschewed: allows to detect ungrammatical and ambiguous structures

$\Rightarrow$ rejection of corpus-based methodologies in linguistics and establishment of a new rationalist orthodoxy

Chomsky's orthodoxy [Hill, 1962]

Chomsky: *The verb 'perform' cannot be used with mass word objects: one can 'perform a task' but one cannot 'perform labour'*

Hatcher: *How do you know, if you don't use a corpus and have not studied the verb 'perform'?*

Chomsky: *How do I know? Because I am a native speaker of the English language*

But, from BNC corpus, it is possible to 'perform magic'!

Going further...

Could Chomsky be wrong?

http://www.timothyjpmason.com/WebPages/LangTeach/CounterChomsky.htm

Chomsky's beef with corpus linguistics

http://www.cty8.com/talandis/categories/gle/Chomsky%27s_beef.htm

Anatomy of a Revolution in the Social Sciences: Chomsky in 1962

http://www.tlg.uci.edu/ opoudjis/Work/KK.html

Other (practical) arguments against the use of corpora

Problem of data processing: pseudo-procedure [Abercrombie, 1971]

No computer, hence processing 11 million words with human eyes only [Käding, 1897] is:

- expensive
- time-consuming
- error prone

Exceptions: the proof of the pudding is in the eating

- Phonology
- Acquisition
- Languages variation (regional expressions, sociolects, register)

Also:

- Data can be observed and checked
- Usefulness of the frequency measure (unavailable from introspection), as for POS taggers

Exceptions: even when introspection does not fail

- Failing to find some sentences or grammatical constructions in a corpus may also be an interesting comment on their frequency
- Introspection lacks systematicity (mistakes can also occur during introspection)
- "*Corpus is a more powerful methodology from the point of view of the scientific method*" [Leech, 1992]

Conclusion on the debate

Criticisms allowed to reflect on what a corpus should be and how to work with it. Most important, it was realised that the corpus and the linguist's intuition were complementary, not antagonistic:

"I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore... [but] every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way. My conclusion is that the two types of linguists need one another." [Fillmore, 1992]

At the same time.. far away

- Father Roberto Busa (1949-1967) was the first person to produce a machine-readable corpus: he asked IBM for help to build the first computer concordancer on a huge corpus of medieval philosophy of 10,600,000 words and 5,000,000 more on some less common languages (German, Russian, Aramaic)
- Alphonse Juilland (1956-1970) said he was doing **mechanolinguistics** and developped a very modern methodology: machine-readable corpora, sampling techniques, texts from different genres, different authors, dispersion statistics, even annotation

$\rightarrow$ they set up the foundations of modern corpus linguistics, but did not work on English...

Work arising from the study of English grammar

- Quirk (1960), Survey of English Usage (SEU)
- Francis and Kucera (1961), Brown corpus

→ great influence on corpus linguistics, both in terms of resources (London-Lund corpus, British National Corpus, Lancaster-Oslo-Bergen corpus, etc.) and linguists (Leech).

neo-Firthians

- From J.R. Firth (1957): social context and social purpose of communication are paramount
- Collocations:

> *"You shall know a word by the company it keeps."*

→ great influence in the UK (Sinclair), esp. with COBUILD project (1980).

⇒ two schools for English.

From pseudo-procedure to viable methodology

In the meantime, the linking between the corpus and the computer was completed.

$\Rightarrow$ corpus studies boomed from the 1980s.

The revival of corpus linguistics [Johansson, 1991]

Date	Nb of Studies
To 1965	10
1966-1970	20
1971-1975	30
1976-1980	80
1981-1985	160
1985-1991	320

Today

- More rigorous corpus linguistics
- (still) Mostly textual corpus linguistics
- Mostly in English
- Boom since the 80s

Tomorrow? [Church, 2011]

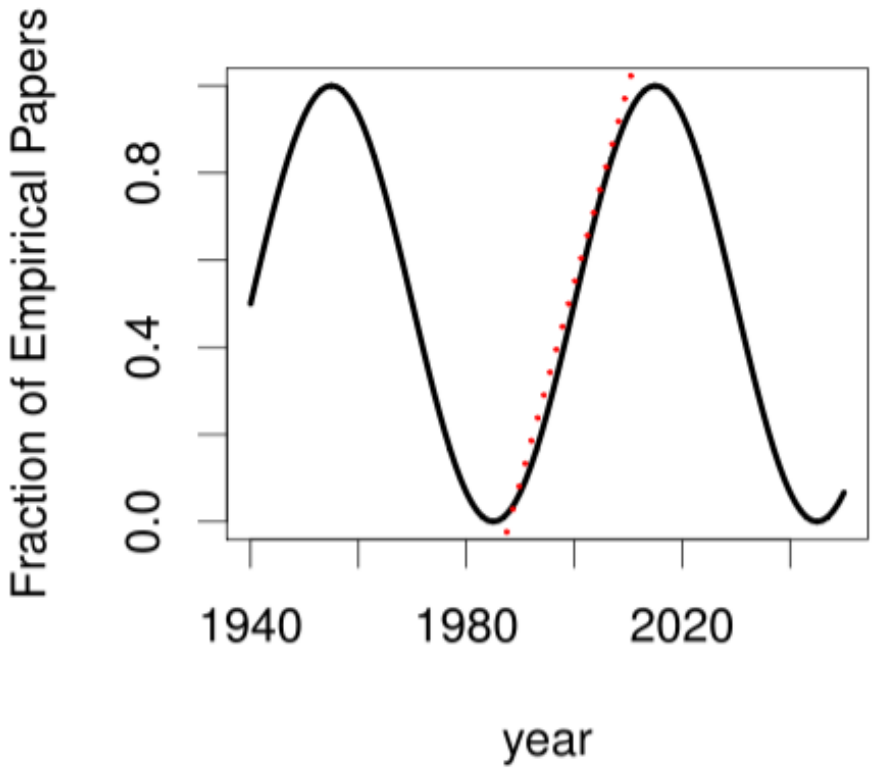


FIGURE 2 An extreme view of the literature, where the trend in Figure 1 (denoted by a dashed red line) is dominated by the larger oscillation every couple of decades. Note that that line is fit to empirical data, unlike the oscillation which is drawn to make a point.

- competence vs performance
- conclusions of the debate
- self-test

- Read carefully: [Sampson, 2000]
  (http://rsta.royalsocietypublishing.org/content/358/1769/1339.full.pdf+html)
- In what does it agree with what we just saw?
- In what does it contradict what we just saw?


Abercrombie, D. (1971).
Studies in phonetics and linguistics.
Oxford University Press, London.


Chomsky, N. (1965).
Aspects of the Theory of Syntax.
MIT, Cambridge, MA.


Church, K. (2011).
A pendulum swung too far.
Linguistic Issues in Language Technology - LiLT, 6.


Fillmore, C. (1992).
"'corpus linguistics' vs. 'computer-aided armchair linguistics'".
In Directions in Corpus LinguisticsProceedings from a 1992 Nobel Symposium on Corpus Linguistics, Stockholm, pages 35–60. Mouton de Gruyter.


Käding, F. W. (1897).

Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch Arbeitsausschüss der deutschen Stenographie-System. Steglitz bei Berlinn: Selbstverlag.


Leech, G. (1991).
The state of the art in corpus linguistics.
English Corpus Linguistics: Linguistic Studies in Honour of Jan Svartvik, pages 8–29.


Leech, G. (1992).
Corpora and theories of linguistic performance.
In Svartvik, J., editor, Directions in corpus linguistics: proceedings of Nobel symposium 82, pages 125–148, Berlin and New York. Mouton de Gruyter.


Morrill, G. (2000).
Incremental processing and acceptability.
Computational Linguistics, 26, 3:319–338.


Rastier, F. (2004).
Enjeux épistémologiques de la linguistique de corpus.

In Texto !

Sampson, G. (2000).
The role of taxonomy in language engineering.
Philosophical Transactions of the Royal Society of London. Series A:Mathematical, Physical and Engineering Sciences, 358(1769):1339–1355.