Corpus Linguistics: corpora

Karën Fort
karen.fort@inist.fr

November 18, 2011

Introduction

Which view on corpora?

Well-known projects

(on) Corpus Linguistics?

Representativeness, Balance and Sampling

Sources

Most of this course is largely inspired by:

- Corpus Linguistics [McEnery and Wilson, 1996],
- Cédrick Fairon's and Anne Catherine Simon's (Université de Louvain) course: Méthodologie de l'analyse de corpus en linguistique.

Corpus definition (reminder)

A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic [and/or extra-linguistic] criteria in order to be used as a sample of the language

[Sinclair, 1996]

Introduction

Which view on corpora?

Well-known projects

(on) Corpus Linguistics?

Representativeness, Balance and Sampling

?

text

?

text speech

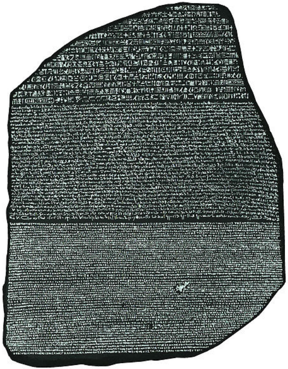
?

text speech music

?

text speech music video

Monolingual / Multilingual



Monolingual / Multilingual

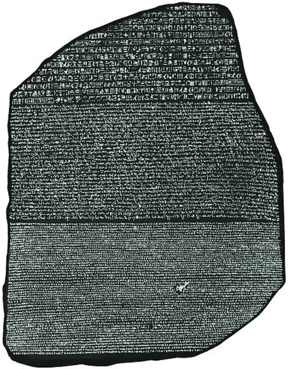
"The Rosetta Stone is a fragment of an Ancient Egyptian granodiorite stele, the engraved text of which provided the key to the modern understanding of Egyptian hieroglyphs. The inscription records a decree that was issued at Memphis in 196 BC on behalf of King Ptolemy V. The decree appears in three texts: the upper one is in ancient **Egyptian hieroglyphs**, the middle one in **Egyptian demotic script**, and the lower text in **ancient Greek**."

(Wikipedia, 27th of Nov. 2010)

Monolingual / Multilingual

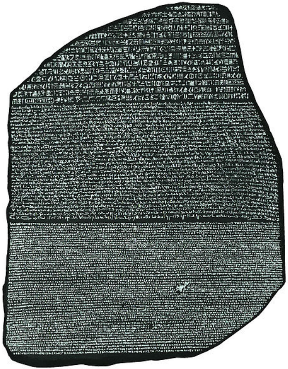
aligned vs comparable

Monolingual / Multilingual



1 or 2 (3) corpora?

Monolingual / Multilingual



1 or 2 (3) corpora? depends on application!

Finite / Open / Dynamic [Baude, 2007]

- Finite (self-contained?): built once and for all as a “complete” corpus [Corpus de référence du français parlé 1; Delic 2004]
- Open: built to integrate new data whether predictively or not [Web, online press]
- Dynamic: sub-category of open corpus, includes Monitor corpus [COBUILD] and Tank corpus [VALIBEL]

Exhaustive / Representative / Balanced / Reference [Baude, 2007]

- Exhaustive: finite corpus containing all the texts for a particular usage (from an author, for example)
- Representative: vague notion, by genres, by sociological sampling, by communication situation
- Balanced: text samples (Brown corpus)
- Reference: built to provide indepth information on a language, big and diverse

Raw data / Constructed object [Baude, 2007]

Natural data vs **created** data (interviews, etc)

Small / Big [Baude, 2007]

What is big?

Organized collection of data / Data bank [Baude, 2007]

Selection?

Bag of words / Texts collection [Baude, 2007]

- **Structured** text or list of *independent* words?
- Complete or partial texts (samples)?

A priori / A posteriori classification [Baude, 2007]

- A priori: extra-linguistic criteria
- A posteriori: internal criteria

Raw / Annotated [Baude, 2007]

Seems obvious, but is **transcription** an annotation?

Short-living / Long-living [Baude, 2007]

- corpus created for **one** research project
- corpus usable in **several** research projects
- corpus with **shareable** annotations (standards)

Conclusion

- Variety of points of view
- Not only texts!

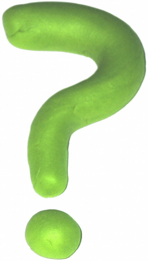
Introduction

Which view on corpora?

Well-known projects

(on) Corpus Linguistics?

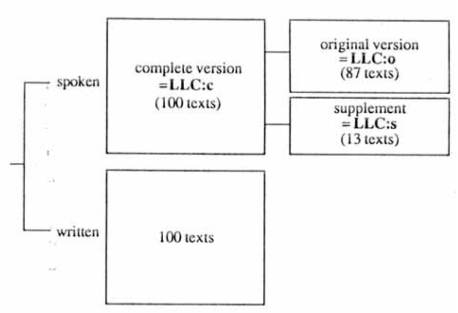
Representativeness, Balance and Sampling



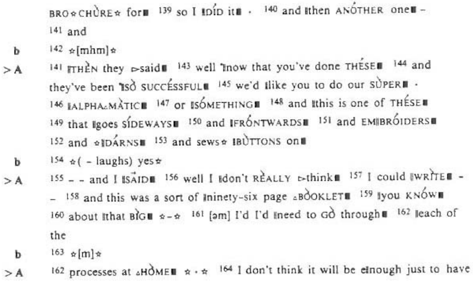
1955-1985: The “Quirk” corpus (aka Survey of English Usage)

- Randolph Quirk
- GB: Survey of English Usage (SEU), University College London
- 1955-1985
- 200 text samples of 5,000 words
- includes 87 spoken texts
- computerized form (500,000 words of spoken British English) known as the London Lund Corpus [Svartvik, 1975]

The computerized SEU corpus (London Lund Corpus)



The London Lund Corpus (reduced transcription)



```
       BRO☆CHÙRE☆ for▮ 139 so I ǀDÍD it▮ · 140 and ǀthen ANÓTHER one▮ –
       141 and
b      142 ☆[mhm]☆
>A     141 ‖THÈN they ▹said▮ 143 well ˈǀnow that you've done THÉSE▮ 144 and
       they've been ˈǀSŎ SUCCÉSSFUL▮ 145 we'd ǀlike you to do our SÙPER▮ ·
       146 ǀALPHA▵MÀTIC▮ 147 or ǀSÓMETHING▮ 148 and ǀthis is one of THÉSE▮
       149 that ‖goes SÍDEWAYS▮ 150 and ǀFRÓNTWARDS▮ 151 and EM‖BRÓIDERS▮
       152 and ☆‖DÁRNS▮ 153 and sews☆ ǀBÙTTONS on▮
b      154 ☆( – laughs) yes☆
>A     155 – – and I ‖SÂID▮ 156 well I ‖don't RÈALLY ▹think▮ 157 I could ‖WRÎTE▮ –
       – 158 and this was a sort of ǀninety-six page ▵BŎOKLET▮ 159 ‖you KNÓW▮
       160 about ǀthat BÎG▮ ☆–☆ 161 [əm] I'd I'd ǀneed to GŎ through▮ 162 ‖each of
       the
b      163 ☆[m]☆
>A     162 processes at ▵HŎME▮ ☆·☆ 164 I don't think it will be eǀnough just to have
```

"[it] retains the following features: tone units (including the subdivision where necessary into subordinate tone units), onsets (the first prominent syllable in a tone unit), location of nuclei, direction of nuclear tones (falls, rises, levels, fall-rises, etc), boosters (ie relative pitch levels), two degrees of pause (brief and unit pauses alone or in combination) and two degrees of stress (normal and heavy). Also indicated are speaker identity, simultaneous talk, contextual comment ('laughs', 'coughs', 'telephone rings', etc) and incomprehensible words (ie where it is uncertain what is said in the recording)."

1961-1979: The Brown Corpus

- *Brown University Standard Corpus of Present-Day American English*
- Francis and Kucera [Kucera and Francis, 1967]: Computational Analysis of Present-Day American English
- US: Brown University, Providence, RI
- 1 million words
- 500 text samples of about 2,000 words each
- publications from 1961
- ready for distribution on magnetic tape in 1964
- tagged in 1979 with TAGGIT [Greene, 1971] with POS, compound forms, contractions, foreign words
- available through NLTK

⇒ let's have a look...

The Brown Corpus with NLTK

```
python
from nltk.corpus import brown
brown.categories()
brown.raw()
brown.words()
brown.sents()
brown.tagged_words()
brown.tagged_sents()
```

The Brown Corpus: some results

- lexicostatistical analysis:
  - American Heritage Dictionary
  - ▸ Zipf's law [Zipf, 1935]
- taggers

The Brown Corpus family

- LOB (Lancaster-Oslo-Bergen corpus of British English, 1978)
- Kolhapur (Indian English, 1978)
- ACE (Australian Corpus of English, also known as the Macquarie corpus, 1986)
- WWC (Wellington Corpus of Written New Zealand English, 1986)
- LCMC (Mandarin Chinese, 1991)

1989: The Penn Treebank 1

- US: UPenn (not free, included in PTB 2, available at LDC)
- one million words (hand-)tagged for part-of-speech:
  - reduced version of the Brown tagset
  - automated (with PARTS) then manual correction, with possibility of multiple tagging
- fully parsed (automatically, then corrected) version of the Brown Corpus
- over 1.6 million words of hand-parsed material from the Dow Jones News Service
  - phrase-structure (bracketed)
  - automated (with Fidditch), then manual correction, with possibility of multiple attachment sites
- used to train the TreeTagger [Schmid, 1997] for English, for example.

1989: The Penn Treebank 2

- US: UPenn (not free, available at LDC)
- includes PTB 1
- new PTB-2 bracketing style, designed to allow the extraction of simple predicate/argument structure
- over one million words of text (1989 Wall Street Journal) provided with this bracketing applied
- annotated text material from the earlier Treebank cleaned up and partly converted

1989: The Penn Treebank 3

- US: UPenn (not free, available at LDC)
- includes part of PTB 2:
  - fully tagged version of the Brown Corpus
  - one million words of 1989 Wall Street Journal
- Switchboard (telephone conversations) tagged, dysfluency-annotated, and parsed text.
- Brown parsed text

1991-1994: The British National Corpus (BNC)

- GB: UCREL (Lancaster University), the British Library and publishers (Oxford University Press)
- not free
- 100 million words
- samples of 45,000 words taken from various parts of single-author texts
- tagged with CLAWS4 (Garside), not corrected, ambiguities kept (error rate evaluated on a 50,000 words sample)
- 10% of spoken corpus
- encoded using TEI (ref. course on Annotations)

The British National Corpus family

- BNC World Edition (enhanced BNC, 2001)
- BNC XML Edition (2007)
- BNC Sampler and BNCBaby (subsets)

1990: The International Corpus of English (ICE)

- initiated by Sydney Greenbaum (SEU)
- set of corpora, some are freely available for research
- 20 research teams from 20 countries
- 20 corpora of 1 million words from 500 texts of 2000 words
- majority of spoken texts (60%)
- (automatic, then corrected) annotations for:
  1. textual markup,
  2. discourse phenomena (false starts, hesitations, etc)
  3. POS tagging and
  4. syntactic parsing (phrase-structure)

Conclusion?

Conclusion

- Big? 1 million words to 100 million words in 30 years!
- Evolution towards speech
- Evolution towards more complex annotations

A **biased** view on corpora

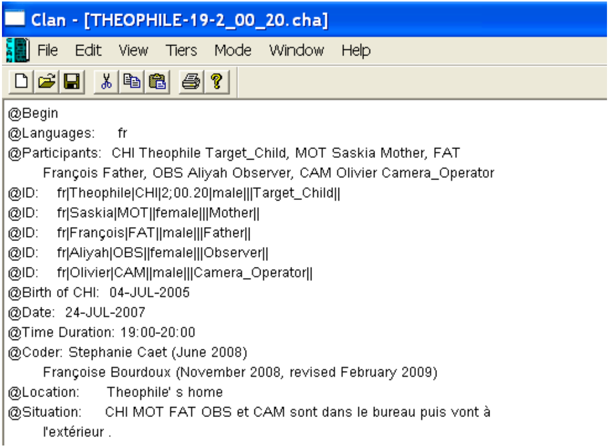
- availability?
- English
- mostly sample-based corpora
- mainly written texts
- general
- annotations quality?

⇒ your presentations should provide us with a larger (if not unbiased) view

1984: The CHILDES corpus

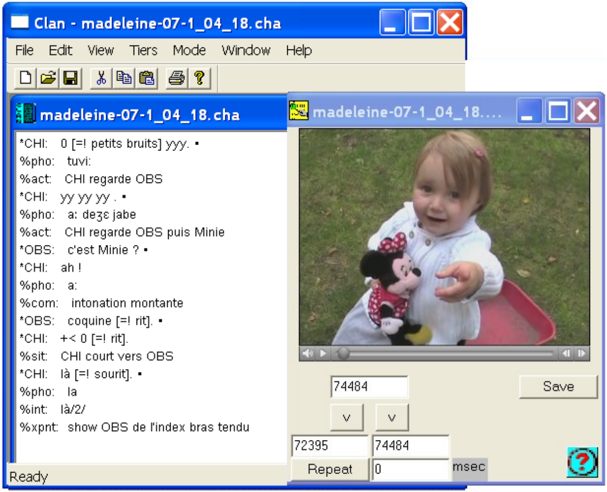
- Child Language Data Exchange System
- US: CMU
- constituted of 3 elements:
    1. CHAT, a transcription and coding format
    2. a database
    3. CLAN, a series of applications allowing to process and analyse data: words, grammar, mistakes, contexts, prosody, accentuation, breaks,...
- freely available

CHILDES: example of heading


Clan - [THEOPHILE-19-2_00_20.cha]
File Edit View Tiers Mode Window Help

```
@Begin
@Languages:     fr
@Participants:  CHI Theophile Target_Child, MOT Saskia Mother, FAT
      François Father, OBS Aliyah Observer, CAM Olivier Camera_Operator
@ID:    fr|Theophile|CHI|2;00.20|male|||Target_Child||
@ID:    fr|Saskia|MOT||female|||Mother||
@ID:    fr|François|FAT||male|||Father||
@ID:    fr|Aliyah|OBS||female|||Observer||
@ID:    fr|Olivier|CAM||male|||Camera_Operator||
@Birth of CHI:  04-JUL-2005
@Date:  24-JUL-2007
@Time Duration: 19:00-20:00
@Coder: Stephanie Caet (June 2008)
      Françoise Bourdoux (November 2008, revised February 2009)
@Location:      Theophile' s home
@Situation:     CHI MOT FAT OBS et CAM sont dans le bureau puis vont à
      l'extérieur .
```

CHILDES: example of transcript

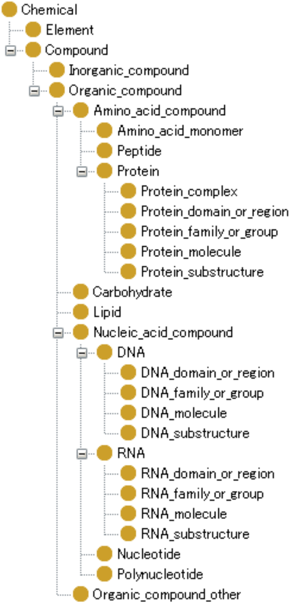


```
*CHI:   0 [=! petits bruits] yyy. ▪
%pho:   tuvi:
%act:   CHI regarde OBS
*CHI:   yy yy yy . ▪
%pho:   a: deʒɛ jabe
%act:   CHI regarde OBS puis Minie
*OBS:   c'est Minie ? ▪
*CHI:   ah !
%pho:   a:
%com:   intonation montante
*OBS:   coquine [=! rit]. ▪
*CHI:   +< 0 [=! rit].
%sit:   CHI court vers OBS
*CHI:   là [=! sourit]. ▪
%pho:   la
%int:   là/2/
%xpnt:  show OBS de l'index bras tendu
```

2003: The GENIA corpus

- Tsujii Laboratory (University of Tokyo)
- 2,000 MEDLINE titles and abstracts (400,000 words) annotated in biology
- annotated manually using an ontology of the domain
- freely available

GENIA ontology



- Chemical
  - Element
  - Compound
    - Inorganic_compound
    - Organic_compound
      - Amino_acid_compound
        - Amino_acid_monomer
        - Peptide
        - Protein
          - Protein_complex
          - Protein_domain_or_region
          - Protein_family_or_group
          - Protein_molecule
          - Protein_substructure
      - Carbohydrate
      - Lipid
      - Nucleic_acid_compound
        - DNA
          - DNA_domain_or_region
          - DNA_family_or_group
          - DNA_molecule
          - DNA_substructure
        - RNA
          - RNA_domain_or_region
          - RNA_family_or_group
          - RNA_molecule
          - RNA_substructure
        - Nucleotide
        - Polynucleotide
      - Organic_compound_other

- used to manually annotate the corpus
- only leaves can be used

A great Web page on corpora!

http://www.lancs.ac.uk/postgrad/xiaoz/papers/corpussurvey.htm

Introduction

Which view on corpora?

Well-known projects

(on) Corpus Linguistics?

Representativeness, Balance and Sampling

Following the debate with Chomsky...

- Data are exploitable by computers
- Data are reliable (at least with a measurable reliability)
  - OK for some automatic annotations (POS tagging)
  - Still pseudo-procedure for other non-annotated corpora (NP recognition)
- Enable searching, sorting, computing. . .

→ Frequencies, ▸ Concordancer

Two stances on corpora

[McEnery and Wilson, 1996]

- collection of authentic computerized texts (including speech transcripts)
- made of sample texts representing a language or a variety of language

[Rastier, 2004]

- structured collection of integral texts
- documented, (potentially) enriched with tags
- put together:
  - in a theoretical way, taking into accounts the genres
  - in a practical way, having an application in mind

Two stances on corpora

[McEnery and Wilson, 1996]

- collection of authentic computerized texts (including speech transcripts)
- made of sample texts representing a language or a variety of language

[Rastier, 2004]

- structured collection of integral texts
- documented, (potentially) enriched with tags
- put together:
  - in a theoretical way, taking into accounts the genres
  - in a practical way, having an application in mind

On which aspects do these definitions differ?

Differences

[McEnery and Wilson, 1996]

- sample texts
- representativeness

→ English “pragmatic“ tradition

[Rastier, 2004]

- integral texts
- structured collection
- documented, tagged

→ French ”philological“ tradition

Corpus-based vs Corpus-driven

[Leech]

- representativeness is considered according to the application
- size is not central
- annotations are usual practice
- studies on lexicons, syntax, pragmatics, semantics, discourse.

⇒ complementary to existing theories

[Sinclair]

- cumulative representativeness (ensured by size)
- the bigger, the better
- annotations are "rejected"
- no disctinction between the different levels of analysis
- holistic approach, collocations (*language patterning*)

⇒ extreme, new paradigm, even new discipline

Introduction

Which view on corpora?

Well-known projects

(on) Corpus Linguistics?

Representativeness, Balance and Sampling

Main References

Should we aim at representativeness?

[McEnery and Wilson, 1996]

- a corpus differs from an archive through representativeness
- necessary condition
- representativeness, sampling and balance are interdependent

[Cappeau and Gadet, 2007]

- we never know what a text is representative of
- demographic representativeness is a question for sociologists, not linguists
- if a speaker[/writer] is representative, of which aspect of his/her personality is s/he representative?

Should we aim at representativeness?

[Rastier, 2004]

No corpus can represent **the** language

⇒ play down the question of representativeness considering it from the specific point of view (vs general) of the application it is collected for

[Rastier, 2004]

"Tout corpus suppose en effet une préconception des applications, fussent-elles simplement documentaires, en vue desquelles il est rassemblé : elle détermine le choix des textes, mais aussi leur mode de "nettoyage", leur codage, leur étiquetage ; enfin, la structuration même du corpus. [...]
... un corpus est adéquat ou non à une tâche en fonction de laquelle on peut déterminer les critères de sa représentativité et de son homogénéité. La linguistique de corpus peut ainsi être objective, mais non objectiviste, puisque tout corpus dépend étroitement du point de vue qui a présidé à sa constitution." [Rastier, 2004]

[Rastier, 2004]

"*Every corpus assumes a detailed knowledge of the application for which it is collected, even if this is a simple documentary application: it not only determines the way texts are selected, but also cleaned up, encoded, tagged and finally the structure of the corpus itself. [...]*
... a corpus is relevant to a task according to which one can determine the criteria for its representativeness and homogeneity. Corpus linguistics can thus be qualified as objective, but not objectivist, as every corpus heavily depends on the point of view that directed its construction." [Rastier, 2004]

How to achieve representativeness?

“Representativeness refers to the extent to which a sample includes the full range of variability in a population.” [Biber, 1993]

⇒ representativeness of a corpus guarantees the generalization of the discoveries made on this corpus to a (variety of) language.

? But how to identify the limits of a “population” to study?

How to achieve representativeness?

- external criteria: different if formal (written style) or informal (oral style):
  - texts genres
  - speech situation
  - demographic characteristics of the speakers
- internal criteria:

  "*The study of corpus words distributions would reveal whether words in a corpus are skewed towards certain varieties and whether in such instances it is accurate to say they are representative of the entire corpus. It would also reflect the stability of the design - whether overall representativeness is very sensitive to particular genres*" (Otlogestwe 2004, quoted in McEnery et al. 2006: 14)

How to achieve balance?

What is the proportion of each type of texts in use in a specific linguistic community?

- balance the representatives of each types of texts (based on a typology of genres)
- balance according to the diffusion/reception of the texts
- balance according to the production of the texts

→ there is **no** valid scientific measure to check the balance of texts in a corpus.

The BNC

- sample: composed of text samples no longer than 45,000 words.
- synchronic: the corpus includes imaginative texts from 1960, informative texts from 1975.
- general: not specifically restricted to any particular subject field, register or genre.
- monolingual British English: comprises text samples which are the product of speakers of British English.
- mixed: contains examples of both spoken and written language.

Balance in the BNC

Text type	Texts	Percent
Spoken demographic	153	10.08
Spoken context-governed	757	7.07
All Spoken	910	17.78
Written books and periodicals	2688	72.75
Written-to-be-spoken	35	1.98
Written miscellaneous	2688	8.09
All Written	2688	82.82

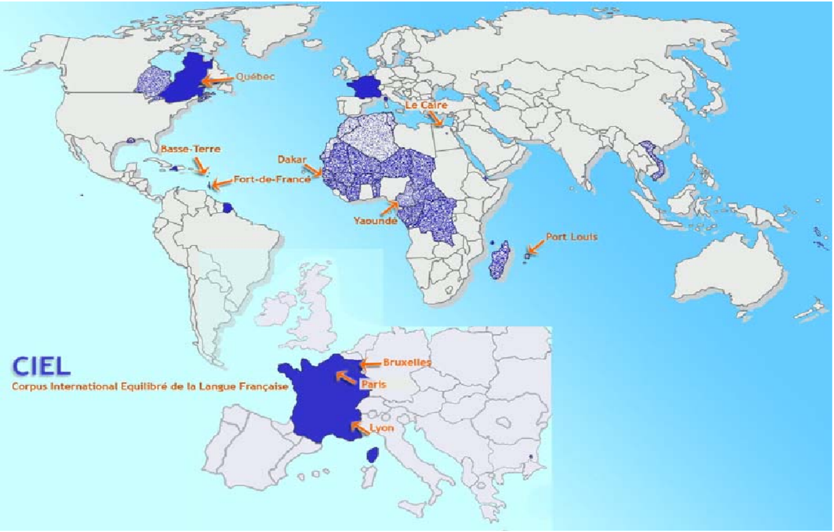
More details...

Balance in the written BNC

Domain	Texts
Applied science	370
Arts	261
Belief and thought	146
Commerce and finance	295
Imaginative	477
Leisure	438
Natural and pure science	146
Social science	527
World affairs	484

More details...

Balance in the CIEL corpus



How to sample?

- Language is infinite (Chomsky)
- The corpus is a sample of a larger population (reduced version of a given population)
- The corpus is generally made of samples:
  - integral texts
  - parts of texts (English-speaking tradition)

→ Examples?

Sample size

To ensure balance and representativeness ⇒ uniform size of texts selected with the application in mind.

Choice between integral texts or parts of texts according to:

- the method / linguistic conception (application?): linguistics of the “word”, “sentence”, “text”
- pragmatic questions: availability (copyright)

Sample size

[Biber, 1993], frequent linguistic phenomena show a stable distribution
⇒ samples of 2,000 words, balanced according to the internal structure of the texts (beginning, middle, end)

Conclusion

- **No** ready-to-use solution to create a reprensentative and balanced corpus
- Importance of documentation
- Keep the application in mind!

- Main projects (SEU, Brown, Penn Treebank, BNC)
- Corpus-driven vs corpus-oriented
- Representativeness and balance depend on the application [Rastier, 2004]

For next course

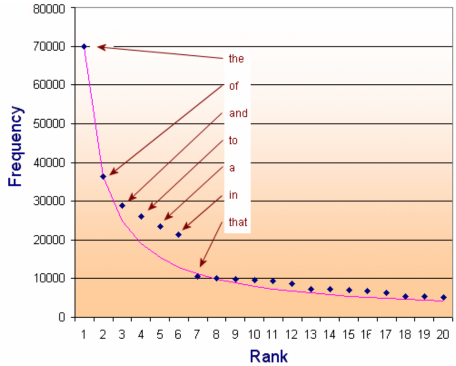
1. Bring your laptop
2. with Transcriber installed
3. and bring earphones!

More about Zipf's law

- given some corpus of natural language utterances, the frequency of any word is **inversely proportional** to its rank in the frequency table.
- ex.: "the" constitutes nearly 7% of the Brown Corpus while about half the total vocabulary of about 50,000 words are hapax legomena.
- Only 135 vocabulary items are needed to account for half the Brown Corpus

Rank	Word	Frequency
1	the	69970
2	of	36410
3	and	28854
20	I	5180

Zipf’s law on the Brown corpus

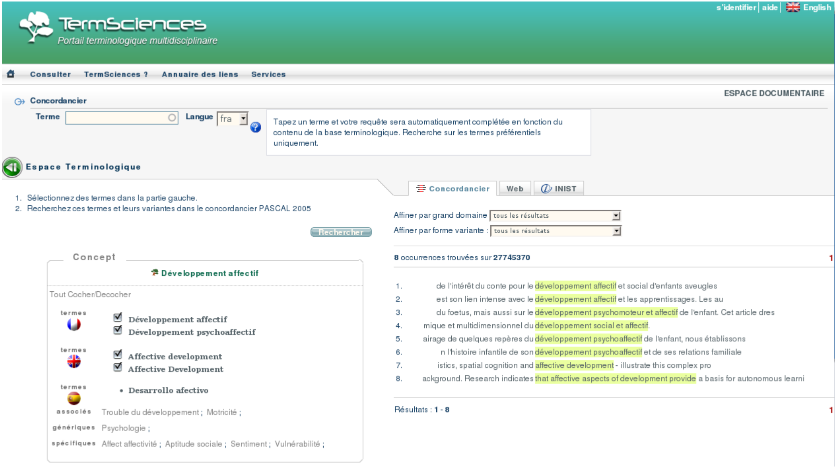


Zipf's law and Language Computation

Read (yes, now!): Introduction of section 2 and section 3 of *Romantics and Revolutionaries* [Steedman, 2011]

▸ Back to Brown

Ex. of concordancer: FastKwic on TermSciences at INIST



Tapez un terme et votre requête sera automatiquement complétée en fonction du contenu de la base terminologique. Recherche sur les termes préférentiels uniquement.

1. Sélectionnez des termes dans la partie gauche.
2. Recherchez ces termes et leurs variantes dans le concordancier PASCAL 2005

Développement affectif

Tout Cocher/Decocher

termes
☑ Développement affectif
☑ Développement psychoaffectif

termes
☑ Affective development
☑ Affective Development

termes
- Desarrollo afectivo

associés Trouble du développement ; Motricité ;

génériques Psychologie ;

spécifiques Affect affectivité ; Aptitude sociale ; Sentiment ; Vulnérabilité ;

Affiner par grand domaine tous les résultats

Affiner par forme variante : tous les résultats

8 occurrences trouvées sur 27745370 1

1. de l'intérêt du conte pour le développement affectif et social d'enfants aveugles
2. est son lien intense avec le développement affectif et les apprentissages. Les au
3. du foetus, mais aussi sur le développement psychomoteur et affectif de l'enfant. Cet article dres
4. mique et multidimensionnel du développement social et affectif.
5. airage de quelques repères du développement psychoaffectif de l'enfant, nous établissons
6. n l'histoire infantile de son développement psychoaffectif et de ses relations familiale
7. istics, spatial cognition and affective development - illustrate this complex pro
8. ackground. Research indicates that affective aspects of development provide a basis for autonomous learni

Résultats : 1 - 8 1

Baude, O. (2007).
Contribution des corpus oraux à la linguistique de corpus : une démarche réflexive intégrée.
In Journées de Linguistique de Corpus, Lorient.

Biber, D. (1993).
Representativeness in Corpus Design.
Literary and Linguistic Computing, 8(4):243–257.

Cappeau, P. and Gadet, F. (2007).
L'exploitation sociolinguistique des grands corpus.
Revue française de linguistique appliquée, XII/1:99–110.

Kucera, H. and Francis, W. N. (1967).
Computational Analysis of Present-Day American English.
Brown University Press, Providence, Rhode Island, USA.

McEnery, T. and Wilson, A. (1996).
Corpus linguistics.
Edinburgh University Press.

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

Sinclair, J. (1996).
Preliminary recommendations on corpus typology.
Technical report, Eagles.

Steedman, M. (2011).
Romantics and revolutionaries.
Linguistic Issues in Language Technology, 6(0).