Freie Universität Berlin

Daten und Fakten

Voraussetzungen und Verfahren der computergestützten Sprachanalyse

Annette Gerstenberg

24. August 2015, Sommerschule Romanistik Mainz

Freie Universität Berlin

Unterlagen zu Vortrag (+ Workshop)

http://userpage.fu-berlin.de/textdaten/DatenMZ/DatenMZ.html

Freie Universität Berlin

August 2015: DRV Sommerschule Mainz

Annette Gerstenberg

TEXTDATEN ROMANISTIK

Unterlagen zum Vortrag *Daten und Fakten: Voraussetzungen und Verfahren der computergestützten Sprachanalyse*

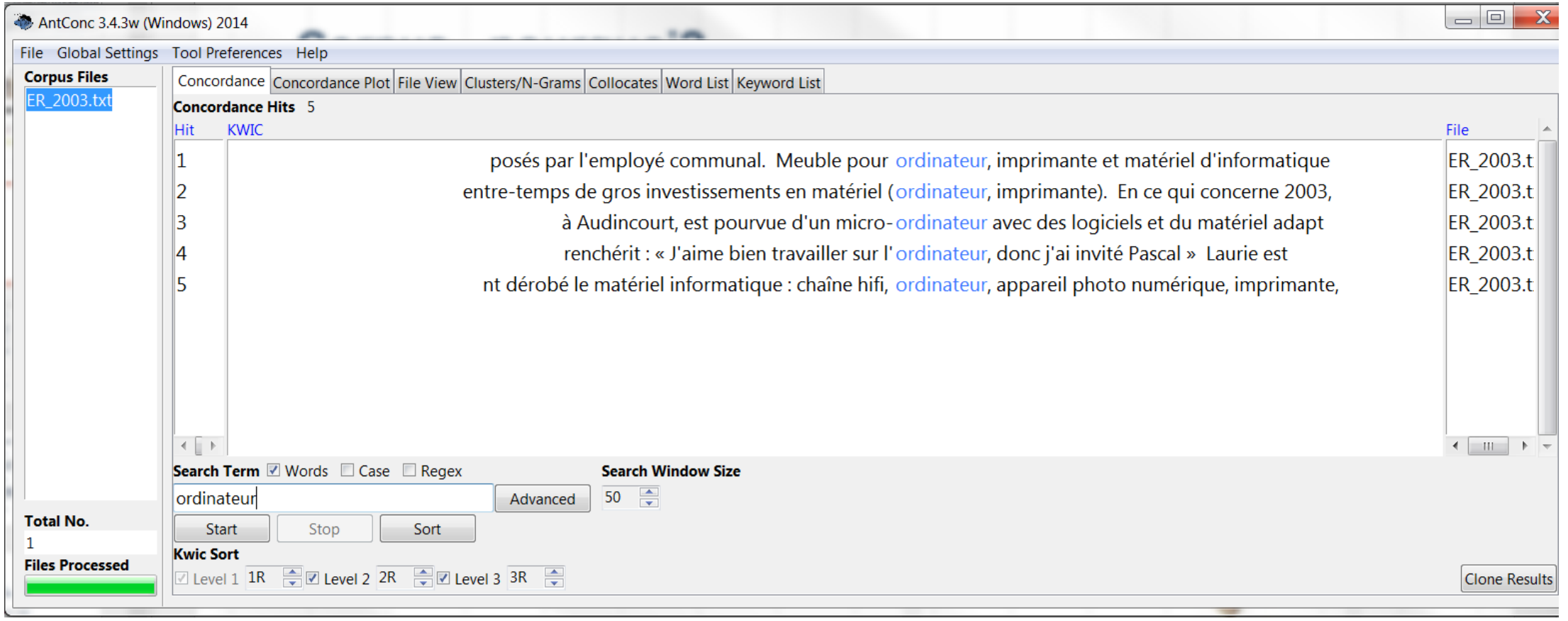
Armchair linguistics does not have a good name in some linguistics circles. A caricature of the armchair linguist is something like this.

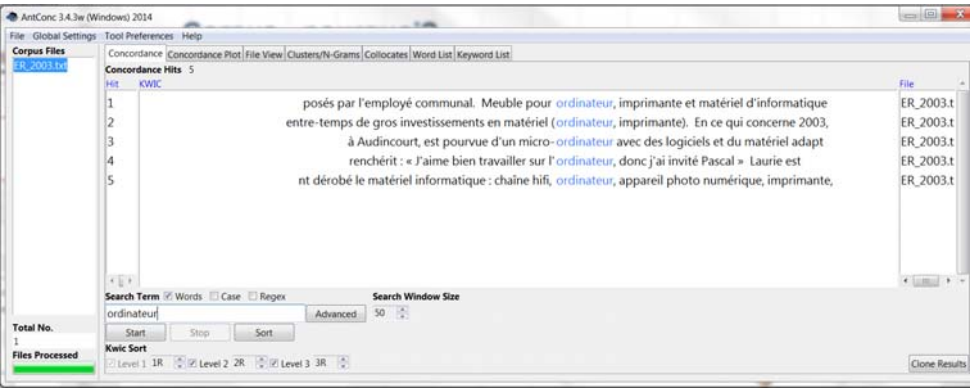
He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head.

Once in a while he opens his eyes, sits up abruptly shouting, ‹Wow, what a neat fact!›, grabs his pencil, and writes something down.

Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. (There isn't anybody exactly like this, but there are some approximations.)

(Fillmore 1992, 35)


AntConc 3.4.3w (Windows) 2014
File Global Settings Tool Preferences Help
Corpus Files
ER_2003.txt
Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List
Concordance Hits 5
Hit KWIC File
1 posés par l'employé communal. Meuble pour ordinateur, imprimante et matériel d'informatique ER_2003.t
2 entre-temps de gros investissements en matériel (ordinateur, imprimante). En ce qui concerne 2003, ER_2003.t
3 à Audincourt, est pourvue d'un micro-ordinateur avec des logiciels et du matériel adapt ER_2003.t
4 renchérit : « J'aime bien travailler sur l'ordinateur, donc j'ai invité Pascal » Laurie est ER_2003.t
5 nt dérobé le matériel informatique : chaîne hifi, ordinateur, appareil photo numérique, imprimante, ER_2003.t
Search Term Words Case Regex
Search Window Size
ordinateur
Advanced
50
Start
Stop
Sort
Kwic Sort
Level 1 1R Level 2 2R Level 3 3R
Clone Results
Total No.
1
Files Processed



Freie Universität Berlin

Corpus linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguist is something like this. He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running word, and he sees his job as that of driving secondary facts from his primary facts.

At the moment he is **busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence**. There isn't anybody exactly like this, but here are some approximations.

(Fillmore 1992, 35)

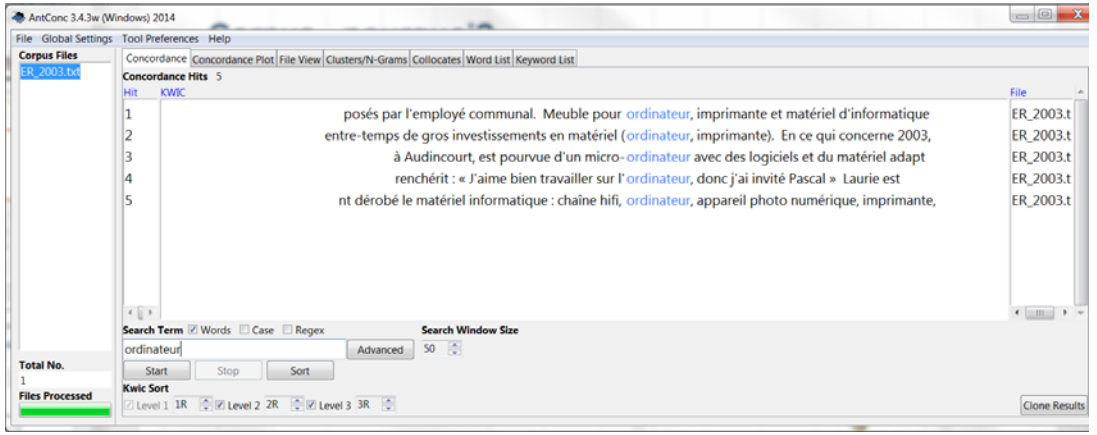
Freie Universität Berlin

My conclusion is that the two kinds of linguists need each other

(Fillmore 1992, 35)

Freie Universität Berlin

+



=

DRV-SOMMERSCHULE 2015 IN MAINZ

Freie Universität Berlin

Daten und Fakten: Programm

1. Voraussetzungen (Texte)
   - Historische Schriftzeugnisse
   - Moderne Schriftzeugnisse
   - Gesprochene Sprache
   - Experimente

2. Verfahren (Corpora)
   - Einführung
   - Daten und Metadaten
   - Basisoperationen
   - Werkzeuge

Freie Universität Berlin

The identification of even the most elementary linguistic datum therefore presupposes an abstraction and a semiotic operation

Lehmann 2004: 183

Freie Universität Berlin

Datentypen nach Arbeitsphase

Originaldaten	Gespräch, Interview, Erzählung…
Rohdaten	Tondatei, ggf. Tonaufbereitung
Primärdaten	Transkript, Basisannotation
Strukturdaten	Linguistisch annotierte Daten

(Himmelmann, Vortrag RUB 2010)

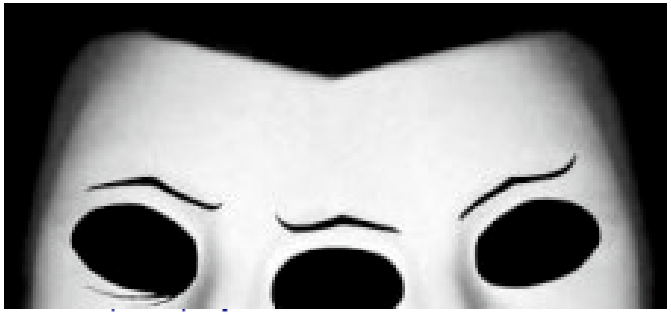
Freie Universität Berlin

datum ? factum ?

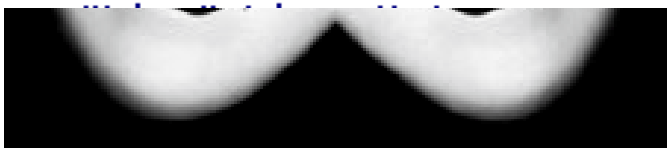
Ziele

Sprache kommt nicht aus der Steckdose.

Mazarin à d'Avaux, Paris, 10 avril 1648
Édition: AE, CP All. 79 fol. 113, fol. 114' = APW II B 8, 114



```
<token pos="KON" lem="parce">parce</token>
<token pos="PRO" lem="que">qu'</token>
<token pos="VER" lem="<unknown>">ilz</token>
<token pos="ADJ" lem="<unknown>">estoyent</token>
```



Hinter der Such-MASKE.

Freie Universität Berlin

1. Voraussetzungen: Texte und Kontexte

- Überblick über unterschiedliche Texte
- Respekt der Entstehensbedingungen
- Aufmerksamkeit für Weichenstellungen
- Beispiele

Freie Universität Berlin

1.1: Historische Texte

- Originalquellen
  - Manuskriptüberlieferung: Es gibt nicht "den" Text
  - Alte Drucke: Ebenfalls graphische Varianz
- Editionen
  - Welche Textzeugnisse werden zu Grunde gelegt?
  - Welche Transkriptionskriterien werden verwendet?

Freie Universität Berlin

1.1: Welche Entscheidungen? Transkriptionskriterien

- Typographie und Layout: Zeilen-, Seitenwechsel; Marginalien
- Grapheme: ‹u›/‹v›
- Graphische Latinismen: ‹pt›, ‹et›
- Wortabstände
- Interpunktion, Satzgrenzen
- Eingriffe in den Text, Umgang mit "Fehlern"

Freie Universität Berlin

1.1: Linguistische Relevanz

- Ein oder mehrere graphische Systeme?
- Welche phonologischen Schlüsse sind zulässig?
- Ist Varianz als graphisch oder als morphologisch einzustufen?
- Erlaubt die Transkription der Interpunktion die syntaktische Analyse?
- Werden "gleiche" Lexeme gefunden?

Freie Universität Berlin

1.1: BSP Nouveau Corpus d'Amsterdam subcorpus id="meun"

vers="oui" ponctuation="non"

mots="25086"

passage="v. 4029-8226/21677"

commentairePhilologique="éd. F. LECOY, ms. BN 1573 (RoSe Jud Tfa 614)"

qualite="cr2"

sourceQualite="DEAF, XG"

commentaireForme="nil"

auteur="GUILLAUME DE LORRIS/JEAN DE MEUNG"

dateComposition="1230ca"

dateManuscrit="1285ca"

lieuComposition="LOIRET"

lieuManuscrit="orl."

sourceDateComposition="DEAF"

sourceDateManuscrit="DEAF"

sourceLieuComposition="DEAF"

sourceLieuManuscrit="DEAF"

genre="poème allégorique sur l'art d'aimer en couplets octosyllabiques"

traditionTextuelle="nil"

analyses="nil"

lignes="4198"

Freie Universität Berlin

1.2: Moderne schriftsprachliche Texte

- Sprachlich normiert (?)
- Mediale Präsentation
- Bedeutung von "Text" in digitalen Medien
- Ausdifferenzierte Textkonventionen

Freie Universität Berlin

1.2: Welche Entscheidungen? Transposition medialer Information

- Integration von Text, Bild, Ton, Video
- Medienspezifische (typo-)graphische Information
- Textkonventionen, Textaufbau

Freie Universität Berlin

1.2: Linguistische Relevanz

- Frequenzeffekte abhängig von Position im Text
- Frequenzeffekte abhängig von Textsortenkonventionen
- Nicht-sprachliche Elemente zum Textverständnis nötig (Semantik und Pragmatik)
- Interne Heterogenität von Diskurstraditionen

1.2: BSP Zeitungskorpus L'Est Républicain

```
<div type="article">
<div>
<head> Second semestre en hausse </head>
<p> A vrai dire, lors du premier semestre alors que les deux maternités, privée et
                    hospitalière étaient encore distinctes, 428 naissances avaient été enregistrées
                    à Bar-le-Duc. </p>
<p> Or depuis le 1er juillet, date de la réunion des deux établissements, 436
                    bébés sont nés dans le service obstétrique du pôle santé, un service qui
                    dispose de trente lits dont dix de structure libérale et qui laisse aux futures
                    mamans le choix de leur praticien public ou privé. </p>
<p> D'ailleurs, le mouvement de grève qui affecte actuellement un certain nombre
                    de maternité et qui est lié à la hausse des primes d'assurances des médecins
                    n'a pas touché le pôle santé et c'est en toute sérénité que sont venus au monde
                    le 31 décembre, Anita, 3,350 kg le dernier bébé de l'année 2002 et Enzo et
                    Meggy les premiers nés de 2003. </p>
</div>
```

Freie Universität Berlin

1.3: Gesprochene Sprache

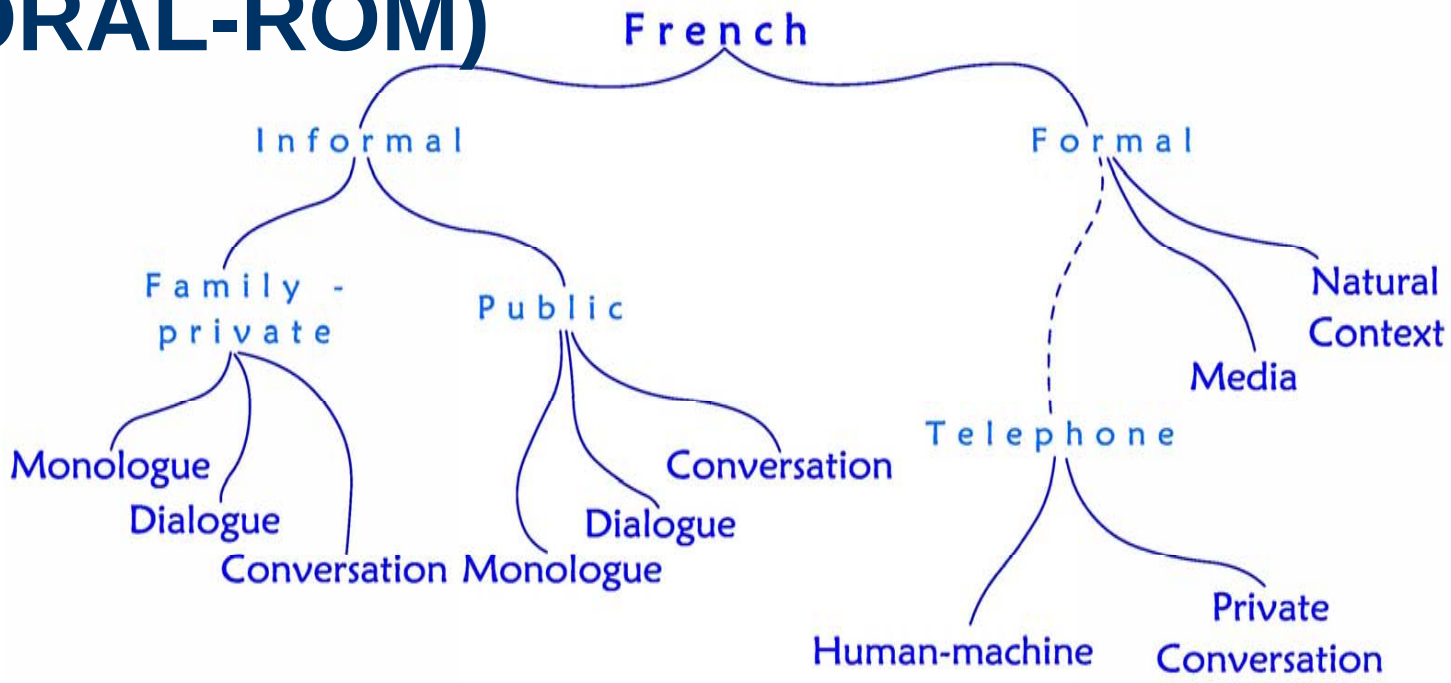
- Nähe-Distanz-Kontinuum
- Situation und Kontext
- Soziolinguistische Faktoren

Freie Universität Berlin

1.3: Welche Entscheidungen?

- Wahl der Dokumentationform: (Protokoll), Audio, Video
- "Authentisches" Setting; Elizitation
- Operationalisierung der soziolinguistischen Indikatoren
- Operationalisierung der pragmatischen Indikatoren

1.3: BSP Texttypen gesprochener Sprache (C-ORAL-ROM)



DELIC

CORAL-ROM

1.3: BSP Variablen der gesprochenen Sprache. Korrelation

- Soziolinguistisch (unabhängig)
  - Alter
  - Geschlecht
  - Sozioprofessionelle Gruppe
- Sprachlich (abhängig): Variable und Varianten
  - /r/ [r] vs. [l] (?) vs. [ʀ]
  - Liaison [pɛʀsɔnzaʒe] vs. [pɛʀsɔnaʒe]

Freie Universität Berlin

1.4: Experimentelle Settings

- Übungen, Aufgaben (SLA)
- Befragungen (Fragebögen, Einschätzung)
- Psycholinguistische Settings (Bildgebung, Eye-Tracking)

1.4: Welche Entscheidungen werden getroffen?

- Wer ist ein native speaker?
- System und Performanz
- Messbarkeit: Auswahl der Skalen

Freie Universität Berlin

1.4: Linguistische Relevanz

- Entweder-oder ? Mehr oder weniger ?

1.4: BSP Introspektion (Grammatikalität)

Abbildung 1: Eine Sieben-Punkt-Skala zur Erhebung von relativen Urteilen

- 25+ Informanten
- 10+ Lexikalisierungen
- Urteile in numerischer Form
- kontrolliertes linguistisches Experimentmaterial

Featherston 2006: 4

Freie Universität Berlin

1.4: Skalen

- Nicht-metrische bzw. kategoriale Skalen
  - Nominalskaliert, z.B. männlich|weiblich
  - Ordinalskaliert, z.B. niedrig|mittel|hoch
- Metrische Skalen
  - Intervallskala, kein natürlicher Nullpunkt, z.B. Jahreszahlen
  - Verhältnisskala, Nullpunkt, z.B. Alter, Prozentzahlen

Freie Universität Berlin

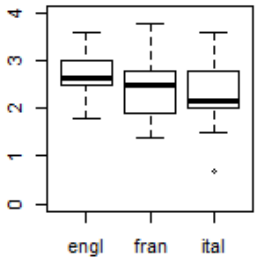
1.4: Wahl der Skala

- Können die Daten in einer Rangreihe angeordnet werden?
  - Nein: Nominalskala
- Ja -> Können Sie bestimmen, wie groß Rangunterschiede sind?
  - Nein: Ordinalskala
- Ja -> Haben die Werte einen natürlichen Nullpunkt?
  - Nein: Intervallskala
    Bsp.: Temperatur in ° Celsius; 10° vs. 20° ≠ "doppelt so warm"
- Skalen mit natürlichem Nullpunkt
  - Ja, mit natürlichen Zählintervallen -> Absolutskala
    Bsp.: Anzahl der Schüler
  - Ja, ohne natürliche Zählintervalle -> Verhältnisskala
    Bsp.: Alter, "doppelt so alt" ist möglich

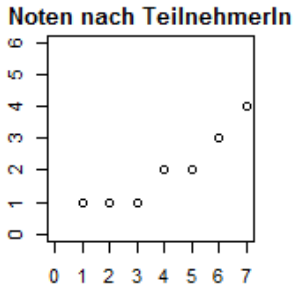
Freie Universität Berlin

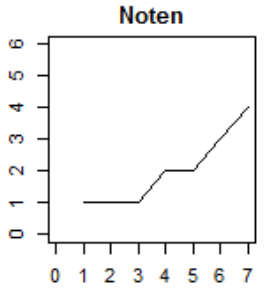
1.4: Graphik, stetig vs. diskret

Boxplot

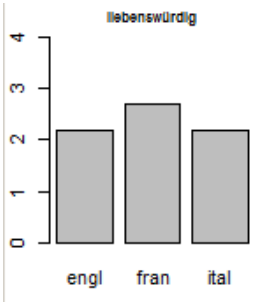


Plot

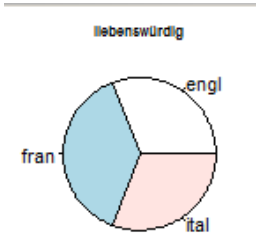




Säulen



Kreisdiagramm



Freie Universität Berlin

Whatever strategy you use, make it defensible, logical and workable

Tagliamonte 2006: 33

Freie Universität Berlin

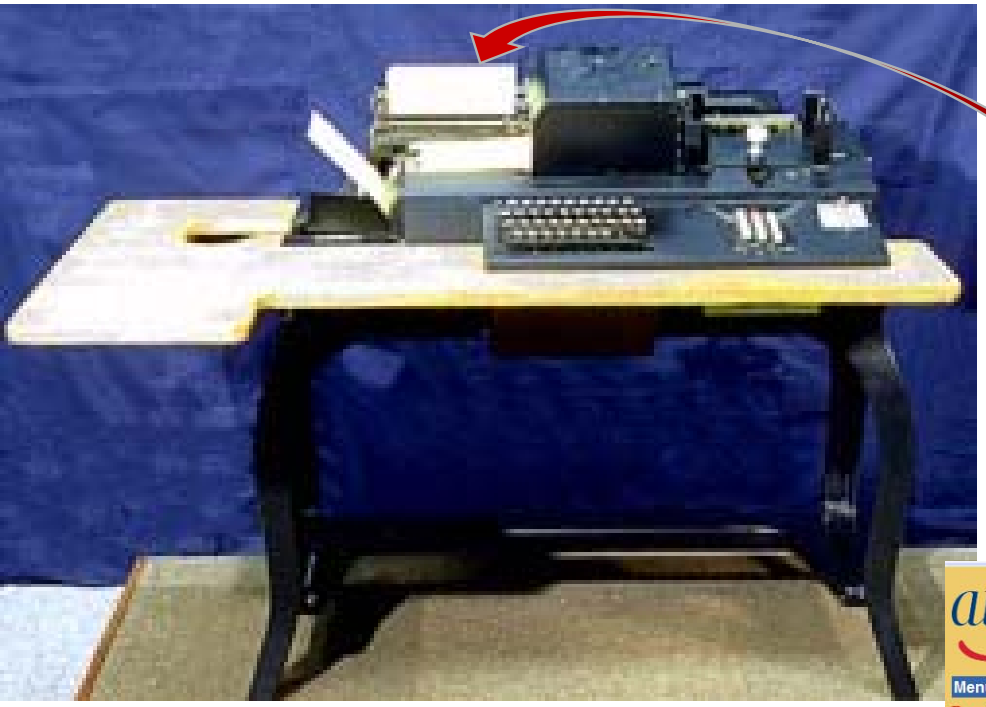
2. Verfahren: Korpuslinguistik

- Einführung
- Daten und Metadaten
- Basisoperationen
- Werkzeuge

Freie Universität Berlin

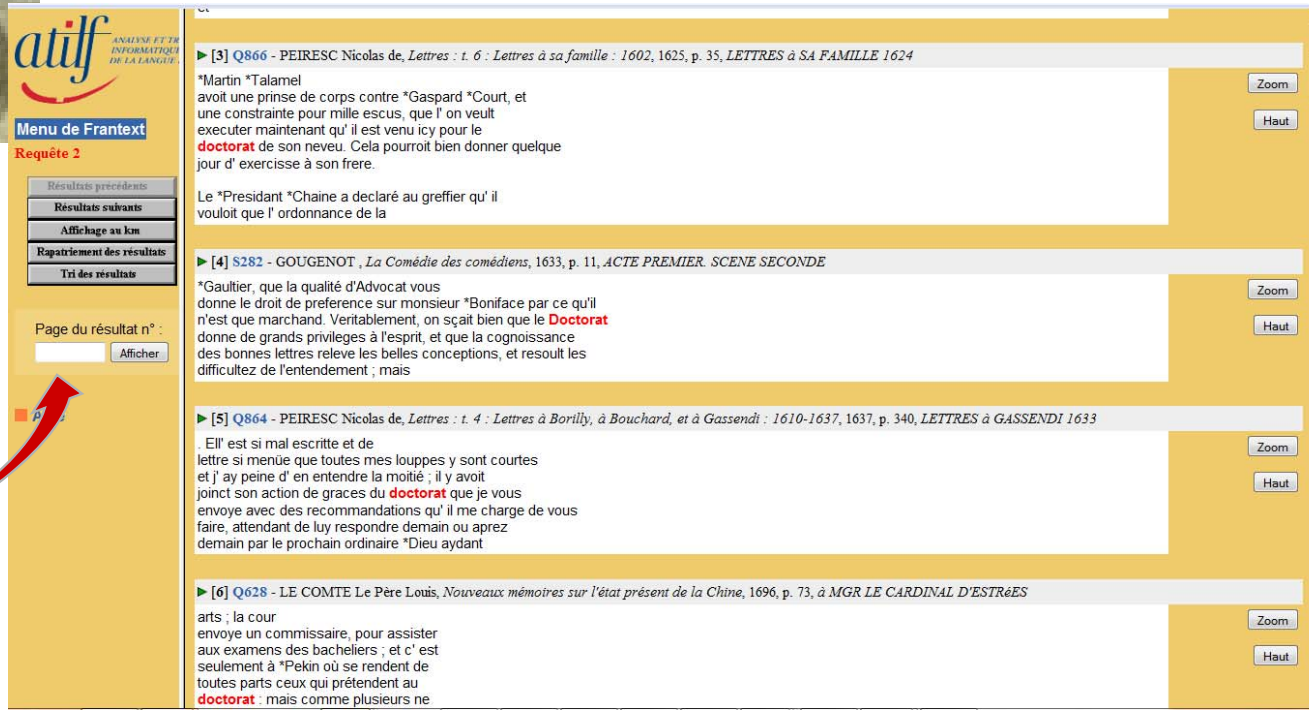
2.1: Korpora und Lexikographie

- Seit den 1960er Jahren systematische Nutzung digitaler Textdaten (Frankreich)
- *Trésor de la langue française* (TLF): literatursprachliches Korpus
  - Textdatenbank seit 1963 (Lochkarten)
  - Textdatenbank Frantext seit 1985 (Internet)

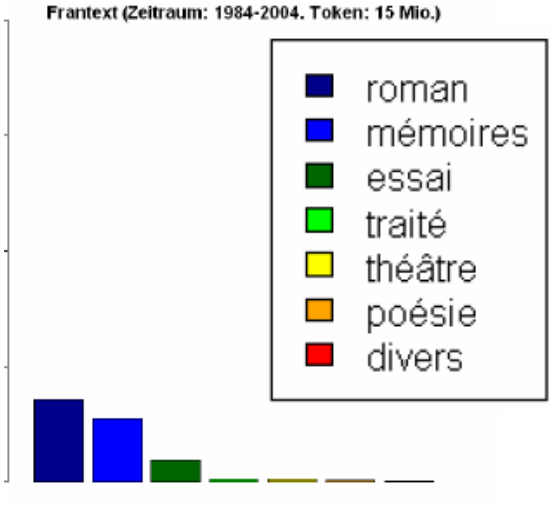


Lochkarte

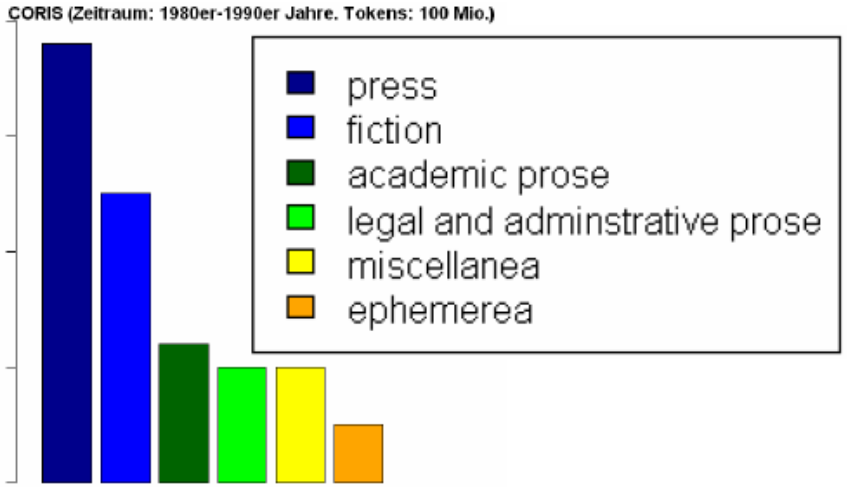
Suchschlitz



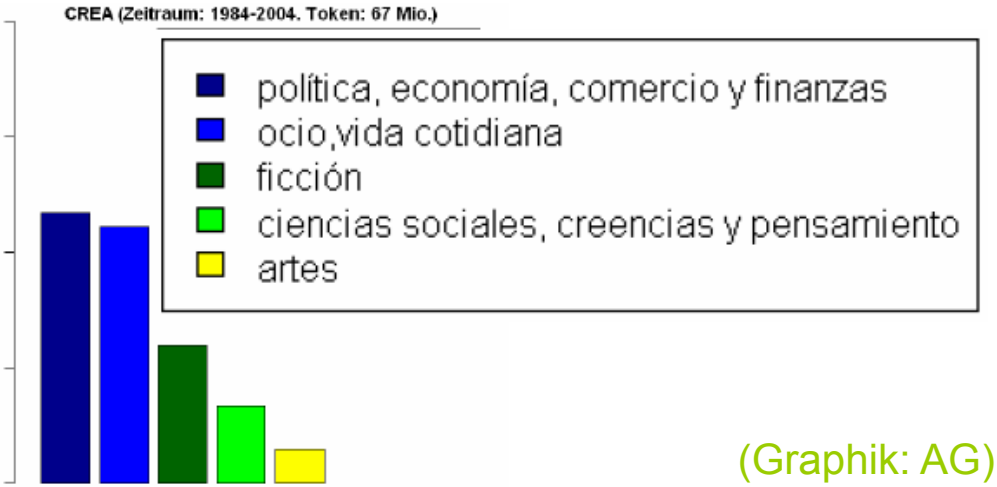
Frantext



CORIS



CREA



(Graphik: AG)

2.1: Definition *Korpus*

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research

(Sinclair 2005)

2.1: Korpuserstellung, Repräsentativität

1. the mode of the text; whether the language originates in speech or writing, or perhaps nowadays in electronic mode
2. the type of text; for example if written, whether a book, a journal, a notice or a letter
3. the domain of the text; for example whether academic or popular
4. the language or languages or language varieties of the corpus
5. the location of the texts; for example (the English of) UK or Australia
6. the date of the texts

(Sinclair 2005)

Freie Universität Berlin

2.1: Grundbegriffe

- Token: Alle graphischen Wörter eines Korpus
- Type: Alle unterschiedlichen Wörter (Lemmata?) eines Korpus
- Frequenzlisten: Alphabetisch oder nach Frequenz geordnet
- KWIC "key word in context" / Konkordanz

2.1: Grundbegriffe, Tokens und Types

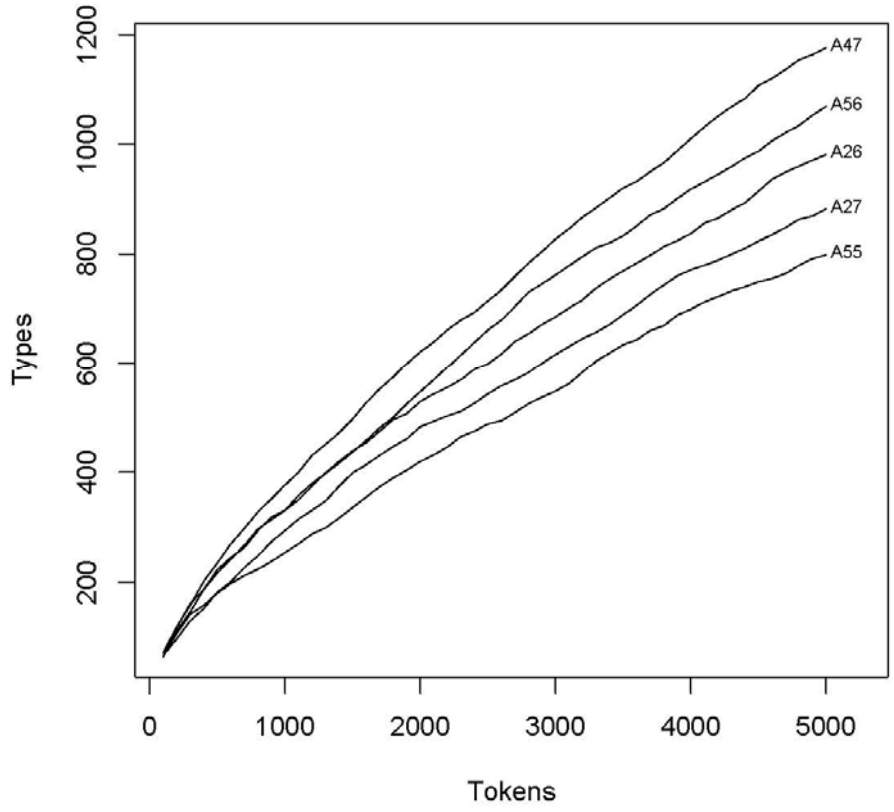
IMPÉRATIF FRANÇAIS @LANGUE_MOLIERE · 15 mars

Albert Jacquard et l'identité nationale.

"Un **Français** ? un **homme** qui s'interroge sur l'**Homme**, en **français**."

(Twitter 2015)

2.1: Grundbegriffe, Type-Token-Relation



(Gerstenberg 2011)

Freie Universität Berlin

2.1: Grundbegriffe, Frequenzlisten

Nach Alphabet

1 ab
4 abandon
1 abandonne
1 abandonner
1 abandonné
1 abattements
1 abattu
2 abe
1 abîmée

Nach Häufigkeit

6692 est
6515 de
4774 c'
4464 et
3848 la
3689 il
3675 le
3564 que
3439 on

Freie Universität Berlin

2.1: Grundebegriffe
KWIC / Konkordanz für frz. *thèse*

bien que cette	**thèse** -là elle est restée dans notre domaine
quand tu as une	**thèse** xxx ça va quoi tu peux trouver
il y en a qui ont plusieurs	**thèse** s en plus
accablants pour confirmer ma	**thèse** à l' heure du crime d' une
qui convient fort bien à la	**thèse** de l'/accusation mais n‘
quant à lui rejeter cette	**thèse** en se raccrochant à un autre
utilisé pour justifier les	**thèse** s néolibérales face à l' impuissance

2.1: Grundbegriffe, Zählbarkeit

- Cleaning, z.B. Links in Internetkorpora
- Tokenization: was ist ein Wort?
  Einfache graphische Definition problematisch
  - Apostroph, aber: *aujourd'hui*
  - Leerzeichen, aber: *in fronte a*
  - Bindestrich im Wortinnern, aber: *vas-y*

Freie Universität Berlin

2.2: Daten und Metadaten

- "Text"
- Linguistische Information
- Sprachexterne Information

2.2.: Metadaten C-ORAL-ROM

@Title: Plongée en Mer Rouge $: ffammn01
@Participant PIE, x
@sex: man
@age: 41-60
@edu: primary school or illiteracy
@prof: employee for Renault
@int: interviewed
@loc: Lyon
@Date: 061999
@Place: Saint Symphorien d'Ozon
@Situation: interview in a kitchen
@Topic: PIE tells his passion for deep-sea diving
@Source: CRFP
@Class: informal, family_private, monologue
@Length: 24'08"
@Words: 4992
@Acoustic_quality: A
@Transcriber: Vanessa BAGUET, Stéphanie BONAL
@Revisor: Mylène BLASCO-DULBECCO, Magali SEIJIDO
@Comments: Presence of other people

Freie Universität Berlin

2.2: Markupsprache

- Trennung von "Text" und Information
- Robuste Datenkodierung
- Beispiel XML "e**X**tensible **M**arkup **L**anguage"
  - Wert-Attribut-Paare in "Tags"
  - Standards z.B. nach TEI (Text Encoding Initiative)

Freie Universität Berlin

2.2: XML-Annotation (NCA)
Attribut (z.B. lemma) Wert (z.B. "et")-Paare

```
<s line="1">
<word pos="PON:cit" deespos="" taggerpos="PON:cit" lemma="QUOTES">'</word>
<word pos="CON:coord" deespos="331" taggerpos="CON:coord" lemma="et" src="ST">et</word>
<word pos="PROCON" deespos="600" taggerpos="PROCON" lemma="si" src="C">si</word>
<word pos="PRO:pers:obj:3:masc:sg:unb" deespos="432" taggerpos="PRO:pers" lemma="il" src="CS">l</word>
<word pos="VER:pres:1:sg" deespos="511*" taggerpos="VER" lemma="aidier|avoir" src="+Md|+CMI">ai</word>
<word pos="PRO:pers:suj:1:sg:unb" deespos="411" taggerpos="PRO:pers" lemma="je" src="CIST">je</word>
<word pos="VER:pper:femi:sg" deespos="582" taggerpos="VER" lemma="perdre" src="+IM">perdue</word>
<word pos="NOM:obj:masc:sg" deespos="002" taggerpos="NOM" lemma="espoir" src="+IT">espoir</word>
</s>
```

Freie Universität Berlin

2.3: Basisoperationen

- Lemmatisierung (Grundwort)
- Part-of-Speech (Kategorie)
- Parsing (syntaktische Information)
- Prosodische Annotation
- Semantische Annotation
- Informationsstrukturelle Annotation
- ...

2.3: Part-of-Speech Tagging

Cleaning: Textaufbereitung

- Entfernen aller Informationen, die nicht mitgezählt werden sollen
- Wortgrenzen definieren (Leerzeichen)

Werkzeuge

- TreeTagger
- Cordial (frz.)

Ausgabe aus Cordial

Token	Lemma	Tag
j'	je	Pp1.sn
ai	avoir	Vaip1s
quitté	quitter	Vmpasm

Freie Universität Berlin

2.3: Part-of-Speech in Europarl, French tagset TreeTagger (Achim Stein)

ABR abbreviation

ADJ adjective

ADJ:num numeral

ADV adverb

CON:coo coordinating conjunction

CON:sub subordinating conjunction

DET:def definite article

DET:indef indefinite article

INT interjection

NOM noun

NPR proper noun

PON punctuation PON:comma comma ,

PON:sep sentence-ending punctuation . ! ?

PRE preposition

PRE:det preposition plus article au, du, aux, des

PRO pronoun PRO:demo demonstrative pronoun PRO:indef indefinite pronoun PRO:inter interrogative pronoun PRO:pers personal pronoun PRO:poss possessive pronoun mien, tien, … PRO:rela relative pronoun

SYM symbol

VER:aux auxiliary verb VER:cond verb conditional VER:futu verb futur VER:impe verb imperative VER:impf verb imperfect VER:infi verb infinitive VER:pper verb past participle VER:ppre verb present participle VER:pres verb present VER:simp verb simple past VER:subi verb subjunctive imperfect VER:subp verb subjunctive present

2.3: Tagset PoS C-ORAL-ROM

PoS	Tag	Sub-type	Examples
Adjectives	ADJ:ORD	Ordinal	premier, deuxième, troisième
Adverbs	ADJ:QUA	Qualifying	petit, grand, vrai
	ADV		ne, pas, oui, alors, très, pratiquement
Conjunctions	CON:COO	Coordination	et, o, mais
	CON:SUB	Subordination	que, parce que, comme, quand, si
Determiners	DET:DEF	Definite	le, la, les
	DET:DEM	Demonstrative	ce, cette, ces
	DET:IND	Indefinite	une, une, tout, quelques, plusieurs
	DET:INT	Interrogative	quel
	DET:POS	Possessive	mon, ma, ton, ta
Interjections and discourse particles	INT		ben, bon, hein, mh, ah
Nouns	NOM:COM	Common	heure, temps, travail, langue
	NOM:PRO	Proper	France, Marseille, Freud, Roosevelt
Numerals	NUM		deux, trois, mille, cent
Prepositions	PRE		de, à, pour, dans, sur
Pronouns	PRO:DEM	Demonstrative	ce, ça, celui, cela, ceci
	PRO:IND	Indefinite	un, une, tout, rien, quelqu'un
	PRO:PER	Personal	je, tu, il, elle, y, en, se
	PRO:POS	Possessive	mien
	PRO:RIN	Relative/interrogative	qui, que, où, quoi, dont, laquelle
Verbs	VER:CON:PRE	Conditional, present	aurait, serait, dirais
	VER:IMP:PRE	Imperative, present	attends, allez, écoutez
	VER:IND:FUT	Indicative, future	sera, aura, fera, seront pourra, faudra
	VER:IND:IMP	Indicative, imperfect	était, avait, faisait, fallait, disait, allait
	VER:IND:PAS	Indicative, past	fut, vint
	VER:IND:PRE	Indicative, present	est a ai sont va ont peut fait sais suis
	VER:INF		avoir, parler, faire
	VER:PAR:PAS	Participle, past	fait dit été eu vu pris pu mis
	VER:PAR:PRE	Participle, present	étant disant faisant ayant
	VER:SUB:IMP	Subjunctive, imperfect	fût, vînt
	VER:SUB:PRE	Subjunctive, present	soit ait puisse fasse aille
Uncategorisable	XXX:ETR	Foreign word	check up, Eine Sache
	XXX:EUP	Euphonic particle	-t-, l'
	XXX:TIT	Title	"Tapas_Café", "With_Full_Force", "Retour_des_Vampires", "Fables_de_La_Fontaine"

(C-ORAL-ROM 2005)

Freie Universität Berlin

2.4: Text- und Analysewerkzeuge

- XML-Editoren: Oxygen, JEdit
  - Reguläre Ausdrücke, XPath
- Texteditoren: Notepad++, Textpad, installierte Editoren
  - Reguläre Ausdrücke,
- Konkordanzprogramme: AntConc, SCP
  - Menügesteuert
- Statistik/Korpusanalyse: R
  - Kommandozeilenbasiert (R) oder menügesteuert

Freie Universität Berlin

2.4: Transkriptionswerkzeuge

- Alignierte Transkription: Hinzufügen von Zeitinformation
- Ein Programm für Transkript, Audio, Video
- Unterschiedliche Funktionen
  - PRAAT
  - Transcriber
  - ELAN

Freie Universität Berlin

Unterlagen zu Vortrag (+ Workshop)

http://userpage.fu-berlin.de/textdaten/DatenMZ/DatenMZ.html

Freie Universität Berlin

August 2015: DRV Sommerschule Mainz

Annette Gerstenberg

TEXTDATEN ROMANISTIK

Unterlagen zum Vortrag *Daten und Fakten: Voraussetzungen und Verfahren der computergestützten Sprachanalyse*

Zitierte Texte

Fillmore, Charles. 1992. 'Corpus linguistics' or 'Computer-aided armchair linguistics'. In Jan Svartvik (ed.), *Directions in Corpus Linguistics*: *Proceedings of Nobel Symposium 82 (Stockholm, 4–8 August 1991)*, 35–60. Berlin, New York: de Gruyter.

Gerstenberg, Annette. 2011. *Generation und Sprachprofile im höheren Lebensalter: Untersuchung auf Basis eines Korpus biographischer Interviews*. Frankfurt am Main: Klostermann.

Lehmann, Christian. 2004. Data in Linguistics. *The Linguistic Review*(21). 175–210.

Sinclair, John. 2005. Corpus and Text - Basic Principles. In Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, 1–16. Oxford: Oxbow Books.

Tagliamonte, Sali A. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Weitere Literaturhinweise: Gerstenberg, Annette. 2013. *Arbeitstechniken für Romanisten. Eine Anleitung für den Bereich Linguistik*. Berlin: de Gruyter.

Zitierte Korpora

APWCF	http://wikis.fu-berlin.de/pages/viewpage.action?pageId=594936338
C-ORAL-ROM	Cresti, Emanuela & Massimo Moneglia (eds.) (2005). *C-Oral-Rom: Integrated Reference Corpora for Spoken Romance Languages*. C-ORAL-ROM. Amsterdam: Benjamins.
CORIS	http://corpora.dslo.unibo.it/coris_ita.html
CREA	http://corpus.rae.es/creanet.html
Est Républicain	http://www.cnrtl.fr/corpus/estrepublicain/
Frantext	http://www.frantext.fr/
NCA	http://www.uni-stuttgart.de/lingrom/stein/corpus/#nca