Korpuslinguistik I

Morphologieanalyse und Lexikonaufbau (8. Vorlesung)

Dozent: Gerold Schneider

Übersicht

Literatur:

- [McEnery 96]: Immer noch das klassische Einführungsbuch
- [Biber 98]: Ausführliches Kapitel über Lexikographie (Part I Chapter 2)
- [Kennedy 98]: Gute Übersicht über englische Korpora (Chapter 2)

Was ist Korpuslinguistik?

Korpuslinguistik (im folgenden KL)

- ist Linguistik basierend auf einer grossen Textsammlung=Korpus
- erhielt ihren Namen von einem Instrument der Analyse (wie etwa*Computer*linguistik)
- hat die gesamte Linguistik revolutioniert
- verwendet statistische Methoden über quantitativ-linguistischen Analysen

Korpuslinguistik (KL) und Computerlinguistik (CL)

KL wird von vielen Computerlinguisten nur als Randgebiet der CL aufgefasst, aber:

- Viele CL Anwendungen basieren auf KL
- KL liefert statistische Informationen für CL (z.B. für Disambiguierung)
- KL ist ein Bindeglied zwischen klassischer Linguistik und CL
- Symbolische versus subsymbolische Ansätze (Z.B. KI versus Statistik)
- KL ist prinzipiell auch ohne Computer möglich, obwohl dann extrem aufwendig

Präskriptive versus deskriptive Linguistik

Wörterbücher, Schul- und Fremdsprachengrammatiken werden häufig zu Rate gezogen, um den richtigen Gebrauch der Sprache sicherzustellen. Die Linguistik vergangener Jahrhunderte sah einen ihrer Hauptzwecke darin, den richtigen Gebrauch der Sparche zu kontrollieren: Präskriptive Linguistik.

Die deskriptive Linguistik hingegen beschreibt Sprache so, wie sie wirklich auftritt.

Performanz versus Kompetenz

Siehe [McEnery 96] Chapter 1.

Unter Sprachkompetenz versteht man die Fähigkeit eines Spechers, wohlgeformte Sätze aufgrund von Sprachregeln zu bilden. Chomsky spricht auch von I(nternalisierter) Sprache und versteht darunter ein Regelsystem (z. B. Phrasenstrukturregeln und Transformationen), dass alle grammatischen Sätze generieren kann und alle ungrammatischen Sätze zurückweist.

Unter Performanz versteht man die Summe der Sprachäusserungen, die ein Sprecher einer Sprache von sich gibt. Chomsky spricht auch von E(xternalisierter) Sprache, eine extrem lange Auflistung aller Sätze, die in irgendeinem Kontext je geäussert wurden.

Kompetenz wird auch als Langue, Performanz auch als Parole bezeichnet.

Chomsky sah die Aufgabe der Linguistik vor allem darin, ein psycholinguistisches Modell der Sprachkompetenz zu bauen und verurteilte Performanz als einen untauglichen Spiegel der Kompetenz.

Zwischen Mitte der 50er und den 80er Jahren kam die Forschung in der KL fast zum Aber das Chomskyanische Universalgrammatikmodell liegt immer noch in endloser Ferne liegt und man kann Sprachdaten sinnvoll einsetzen kann

- zur Überprüfung von Hypothesen
- zur Anreicherung von Regeln mit statistischen Wahrscheinlichkeiten, was Ambiguität verringert
- zur detaillierten Sprachbeschreibung, viel zuverlässiger als die Intuition eines Grammatikschreibers
- als Grundlage für eine empirisch-wissenschaftliche Beschreibung der Sprache

→ Sprachwissenschaft an der Schnittstelle zwischen einer empirischen und einer rationalen Wissenschaft.

KL ist das systematische, streng empirische Vorgehen mit Beobachtung und Regelableitung

Der introspektive ("armchair") Linguist geht dagegen in der Art des Rationalismus vor

Anwendungen der Korpuslinguistik

KL, da eher linguistisches Instrument als Teilgebiet der Linguistik, findet Anwendungen in fast allen linguistischen Teilgebieten

Morphologie

- Produktivität: Auf der Basis eines morphologisch annotierten Korpus oder mittels eines zuverlässigen (!) Morphologieanalyseprogramms: Welche Derivationsmorpheme sind produktiv? Welche Partikel- und Derivationsmorpheme treten bei welchen Kernmorphemen auf, welche bei fast allen (produktiv), welche nur sehr eingechränkt?
- Flexion: Welche Flexionsformen treten häufig, welche selten auf (sg/pl,nom/acc, 1./2./3. Person) und gibt es Registerunterschiede? Stimmen unsere Annahmen über Markiertheit und Ikonizität? Werden unzählbare Substantive wirklich nie im Plural verwendet?
- Tagging: Welche Wortart von allen möglichen einer gegebenen Wortform ist häufiger? Und in welchem Kontext? Kann die richtige Wortart aufgrund dieser Angaben immer zugewiesen werden? (siehe Statistisches Tagging und Regelbasiertes Tagging)
- Distribution: Welche Worte und Wortarten sindwie verteilt? Treten gewisse Wortarten in gewissen Sprachen oder Registern häufiger auf?

Lexikologie und Lexigraphie

[Biber 98] Chapter 2 ist ganz diesem Thema gewidmet.

Lexikographen sind schon seit Jahrhunderten bemüht, auf KL abzustützen. Samuel Johnson suchte schon 1755 mühevoll nach Beispielen aus der Literatur für die Einträge in seinem Lexikon. Heute gibt es kaum mehr ein Lexikon, das nicht KL-getestet ist.

KL wird heute verwendet für:

- Semi-automatische Lexikographie
- Semi-automatische Term-Extraktion: Z.B.: Tritt ein best. Wort gehäuft oder nur in einem gewissen Register auf, oder fehlt es in einem allgemeinen Lexikon, so ist es ein Term-Kandidat.

KL-basierte Methoden erlauben zusätzlich lexikologische Angaben zu:

- Sprachgebrauch: Welche deutschen Worte sind die häufigsten (Strukturworte)? Welche Bedeutung eines polysemischen Wortes kommt wie häufig vor?
- Kollokationen: In welchen Kontexten (lokal oder hierarchisch) treten welche Worte auf? Sind diese für Synomyme gleich? Welche grammatischen (Syntax) und semantischen Begleiter (semantische Assoziationen) treten auf?

Syntax

Gegeben sei folgende Trivialgrammatik und ein Parsingaufruf:

```
S -> NP VP.
S -> NP.
NP -> N N.
NP -> N.
VP -> 'walks'.
NP -> 'walks'.
N -> 'John'.

Parse>> John walks.
```

Wie könnte man mit Hilfe eines annotierten Korpus der richtigen Lesart dieses Satzes Vorzug geben?

Häufig werden statistische Gewichtungen zur Disambiguierung von PP-Anbindung oder Adverbskopus verwendet:

```
John sees the man in the park with the telescope.
```

Welche Verben und welche Nomen verwenden welche Präpositionen mit Vorliebe?

```
Heureusement, Jean part pour les vacances.
Jean heureusement survit l'accident.
```

Welche Verben und Adverben (oder Kombinationen) haben eher welchen Skopus?

Semantik

In der Semantik gibt es z.B. Untersuchungen darüber, ob zwei synonymische Ausdrücke tendentiell von den gleichen Worten begleitet werden (Kollokationen). Es gibt auch Versuche, semantische Netzwerke zu konstruieren, bei denen (für referenzierende Worte) die durchschnittliche lokale Nähe im Text als Parameter der semantischen Nähe angenommen wird.

Phonetik

Ein grosser Teil der phonetischen KL untersucht Sprachvariation nach Alter, Dialekt, Ausbildung, Geschlecht etc. Korpora mit gesprochenen Daten sind aber auch interessant da sie noch näher an 'real-world' Sprachgebrauch sind als Textcorpora, also z.B. Grammatik so, wie sie 'in unseren Köpfen' lebt anstatt in Schulbüchern. Dies ist für die Psycholinguistik von Interesse. Viel linguistischen Wissen über Prosodie wurde auch durch Korpusanalysen gewonnen. Grosse Corpora gesprochener Sprache werden in der CL verwendet, um Spracherkennungssysteme möglichst sprecherunabhängig trainieren zu können.

Pragmatik, Diskursanalyse

Meist auch anhand gesprochener Korpora. Wer spricht wann, wie sind die Übergange zwischen Sprechern, wer unterbricht eine Äusserung, wie signalisieren Hörer ihr Interesse an gesprochenen?

Soziolinguistik, Dialektologie

Soziolinguistik ist per definition deskriptive Linguistik. Korpusbasierte Ansätze untersuchen Sprachvariation nach Alter, Dialekt, Ausbildung, Geschlecht und so weiter.

Historische Linguistik

Den meisten qualitativen Veränderungen in der Sprache geht ein langer schleichender soziolinguistischer Prozess voran, in dem eine seltenere Form die üblichere immer stärker vertritt und diese mit der Zeit überflügelt. Nur quantitative Untersuchungen belegen die Präferenzen bei der Auswahl, die letztlich zu einere qualitativen Änderung führen.

Stilistik und Stilometrie

Verschiedene Autoren bedienen sich verschiedener Stilmittel, was sich auch in der quantitativen Sprachbeschreibung niederschlägt. Welcher Autor braucht welche Worte häufiger, macht typischerweise welche Fehler, hat ungefähr welchen Substantiv/Verb Koeffizient etc. Die Stilometrie kann häufig Autoren unbekannter Werke identifizieren oder stellt sich die Frage, ob Shakespeare nur eine Person oder das Pseudonym einer Autorengruppe war.

Grundfragen der Korpuslinguistik

Zusammenstellung des Korpus: Ausgewogenheit, Repräsentativität

Die Auswahl der Texte, die in einen Korpus aufgenommen werden sollen, ist entscheidend. Will man einen genrespezifischen oder einen allgemeinen Korpus zusammenstellen? Welche Texte repräsentieren welches Genre? Schwierig: Wie sollen verschiedene Genres in einem allgemeinen Korpus vertreten sein? Letztlich gibt es prinzipiell nie einen völlig ausgewogenen Korpus.

Statistische Methoden: Aussagekraft, Relevanz

Alleine die Erkenntnis, dass ein gewisses Phänomen in einem Korpus häufiger als in einem anderen vorkommt, macht noch keine statistisch aussagekräftige Aussage. Eine dialektologische Untersuchung, die nur auf einen Probanden jeder Dialektgruppe abstützt, ist statistisch vollkommen wertlos, da die Unterschiede genausogut individueller Präferenzen anstatt dialektaler entspringen können.

Falls der zu untersuchende Häufigkeitsunterschied zwischen zwei (Teil-)Korpora gering ist oder falls nur eine kleine Anzahl Belege gefunden werden konnten, so ist das Risiko gross, dass es sich beim Unterschied nur um zufällige Fluktuationen handelt. Um den Einfluss zufälliger

Schwankungen einschätzen zu können und somit echte von zufälligen Unterschieden trennen zu können, verwendet man eine Reihe von statistischen Tests. Siehe nächste Vorlesung.

Korpusannotation

Ein roher Text erlaubt nur klare Aussagen auf der Wortformenebene. Für gültige Aussagen auf der phonologischen, morphologischen, syntaktischen oder semantischen Ebene darf man sich nicht exklusiv auf fehlerbehaftete automatische Analysemethoden abstützen. Auf der morphologischen Ebene beispielsweise macht auch eine Taggingfehlerrate von nur 2% u.U. eine statistische Aussage völlig zunichte, falls sie teilweise auf die fehlerhaften Tags in einem so getaggten Korpus abstützt.

Deshalb ist es meist notwendig, sich auf zuverlässige, handkorrigierte oder -erstellte Analysen abzustützen, mit denen der Korpus annotiert ist. Eine Annotation auf möglichst vielen Ebenen ist zwar wünschenswert, verteuert und verlängert aber die Zusammenstellung eines Korpus ins Endlose. So beschränkt man sich meist auf wenige Ebenen oder der Korpus bleibt sehr klein. Für viele syntaktisch interessante Aussagen sind die meisten (aufwendig!) syntaktisch annotierten Korpora auch heute noch zu klein.

Als Beispiel einer morphosyntaktischen Annotation siehe STTS aus der ersten Vorlesung.

Verwaltung der Korpusdaten: Speicherung, Zugriff, Visualisierung

Speicherung

Die Verwaltung von grossen Datenmengen stellt Probleme in der Art der Speicherung, die schon auf einen einfachen und schnellen Zugriff vorbereiten soll.

- **Datenbank (DB)**: Effizient, auf schnellen Zugriff ausgelegt. Die Textdaten müssen aber konvertiert werden in ein DB-Format, was aufwendig und für Linguisten nicht unbedingt einfach ist. Die konvertierten Daten sind für Lingusten nicht mehr manuell lesbar.
- **Rohtext:** Sehr flexibel, der Quelltext kann direkt verwendet werden, keine DB-Spezialisten notwendig. Die Speicherung kann aber platzintensiv und der Zugriff langsam sein. Es ist schwierig, so einen Korpora zu handhaben, der nicht vollständig im RAM Platz hat.
- **SGML, XML:** Strukturierter Text, schneller als Rohtext, langsamer als DB, aber manuell lesbar. Ein Kompromiss zwischen DB und Rohtext.
- **proprietäre Formate:** Auch effizient, aber schwierig portierbar und überhaupt nicht mehr manuell lesbar.

Zugriff

- **SQL** (oder andere Datenbankabfragesprache): Der übliche Zugriff auf eine Datenbank. Sehr effizient, ziemlich flexibel.

  Ein SQL-Abfragebefehl hat das folgende Format:

  ```
  SELECT Zu findende Werte
  FROM Tabelle
  WHERE Bedingung
  ```

  Stellen wir uns z.B. folgenden einfachen Korpusaufbau vor:

  ...

SATZ 2112	1	2	3
wort	Ein	Satz	!
tag	DET	NN	PKT

SATZ 2113	1	2	3	4	5	6	7
wort	Der	vorliegende	Text	ist	ein	Beispiel	.
tag	DET	ADJA	NN	VVFIN	DET	NN	PKT

SATZ 2114	1	2	3	4	5
wort	Hier	geht	es	weiter	.
tag	ADV	VVFIN	PRO	PTKVZ	PKT

  ...

Dann würde beispielsweise eine Anfrage nach den Artikeln im 2113. Satz des Korpus folgendes Aussehen haben, und liefert die darunterstehenden Ergbnisse:

```
SELECT wort
FROM SATZ 2113
WHERE tag = DET

> wort = Der
> wort = ein
> ALL RESULTS
```

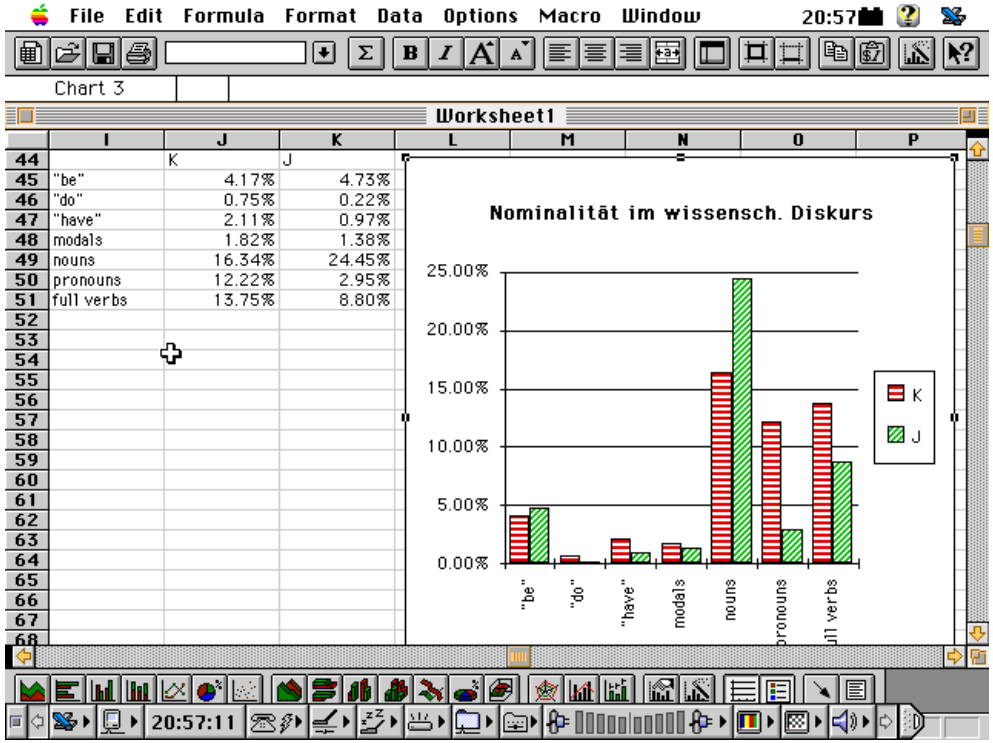
- **Konkordanzprogramme**: Einfach, oft visuelle Ausgabe der Ergebnisse. Die Möglichkeiten sind aber beschränkt und nicht erweiterbar. Siehe ein Beispiel unten.

- **Perl** (oder andere listenoptimierte Programmiersprache): Arbeitet über Rohtext, Listenvariablen oder gar DB-Formate. Eher langsam, aber extrem flexibel.

```
#!local/bin/perl
# Dieses Prgramm sucht STTS Artikel, gibt sie aus und z"ahlt die Types
# (c) GS
while (<>) {                              # lese Korpus wortweise ein
      ($word, $tag) = split(/\_/);       # trenne Wort und Tag bei _
      if ($tag == "DET") {                # ist es ein Artikel (DET)? Falls ja:
            print "$word \n";               # gebe Wort aus
            ++$typecounter{$word};          # erhöhe Zähler für dieses Wort um 1
            }
      }                                  # Ende der while-Schleife: Nimm nächstes Wort
                                         #  und durchlaufe mit ihm die Schleife, etc.
```

Visualisierung

Die extrahierten Daten, in der Regel lange Listen und Tabellen, geben oft noch keinen Überblick über die untersuchten Zusammenhänge. Hier empfiehlt sich der Einsatz eines professionellen Statistikprogrammes oder (für einfachere Aufgaben) des altbekannten MS Excel.



	I	J	K	L	M	N	O	P
44		K	J					
45	"be"	4.17%	4.73%					
46	"do"	0.75%	0.22%					
47	"have"	2.11%	0.97%					
48	modals	1.82%	1.38%					
49	nouns	16.34%	24.45%					
50	pronouns	12.22%	2.95%					
51	full verbs	13.75%	8.80%					

Korpusaufbau

Ein historischer wie thematischer Abriss über den Aufbau der ersten grösseren computerlesbaren Korpora.

Aufteilung der Texte

Schon 1964 begann in den USA ander Brown Universität (daher der Name Brown Corpus) die Zusammenstellung des ersten grösseren computerlesbaren Corpus. Die Zeiten waren schwierig; "it was compiled in the face of massive indifference if not outright hostility from those who espoused the conventional wisdom of the new and increasingly dominant paradigm in US linguistics led by Noam Chomsky" ([Kennedy 98] p. 23). Die damals immense angepeilte Korpusgrösse von 1 Mio Worte wurde innerhalb von 3 Jahren Arbeit erreicht. Im Bestreben um Ausgewogenheit wurden folgende Genres aufgenommen:

Aufteilung des Brown und des LOB Korpus

```
Category A (Press: reportage)
Category B (Press: editorial)
Category C (Press: reviews)
Category D (Religion)
Category E (Skills, trades and hobbies)
Category F (Popular lore)
Category G (Belles lettres, biography, essays)
```

```
Category H (Miscellaneous, mainly Government documents)
Category J (Learned and scientific writings)
Category K (General Fiction)
Category L (Mystery and detective fiction)
Category M (Science fiction)
Category N (Adventure and western fiction)
Category P (Romance and love story)
Category R (Humour)
```

Diesseits des Atlantik standen erst in den 70er Jahren genügend Ressourrcen zur Verfügung um ein ähnliches Projekt in Angriff zu nehmen. Von 1970-78 wurde als gemeinsames Projekt der beteiligten Universitäten der Lancaster-Oslo-Bergen (LOB) Korpus zusammengestellt, nach der gleichen Aufteilung wie der Brown Korpus, ebenfalls mit etwa 1 Mio Worten, womit beispielsweise erstmals ein qunatitativer Vergleich zwischen US und GB Englisch möglich wurde.

1988 folgte ein fast gleich aufgeteilter Korpus für Indisches Englisch, darauf ebenso einer für Australisches, und 1993 einer für Neuseeländisches Englisch. Auch heute noch finden diese Korpora weite Verwendung, nicht zuletzt weil sie klein genug sind um auf heute üblichen PCs bearbeitet zu werden.

Annotierung der Korpora

Schon in den 70er Jahren wurden erste Versuche zum semi-automatischen part-of-speech Tagging vorgenommen. Die Fehlerraten waren aber noch sehr hoch. Erst das für den LOB

Korpus entwickelte CLAWS System erreichte Fehlerraten von unter 5%. Die manuelle Korrektur bleibt aufwendig, im Falle das LOB Korpus erfolgte sie zwischen 1978 und 1983.

Teile diser Korpora liegen auch schon syntaktisch annotiert in hierarchischer Phrasenstruktur vor, aber noch keiner von ihnen vollständig. Parsing und manuelle Korrektur sind noch wesentlich aufwendiger als fürs part-of-speech Tagging.

Heutige Korpora

Heute gibt es viele Spezialkorpora in Englisch, für gesprochenes Englisch, diachrone Sprachentwicklung, Spracherwerb etc. Auch fürs deutsche sind schon einige erhältlich (z.B. der STTS-annotierte Frankfurter Rundschau Korpus mit etwa 250'000 Worten oder der kleine von uns STTS-annotierte Zürcher Universitätskorpus mit 50'000 Worten).

Im Englischen wird die Entwicklung von extrem grossen Corpora vorangetrieben. Der aus dem Cobuild-Projekt entstandene, allerdings ungeaggte Bank of English Korpus umfasst etwa 300 Mio Worte. Der British National Corpus (BNC) ist ähnlich wie Brown oder LOB getaggt und umfasst 100 Mio Worte, aber es ist fraglich, wieviele Taggingfehler die gezwungenermasse oberflächliche Korrektur des automatischen Tagging noch überlebt haben. Am Englischen Seminar der Uni Zürich ist Prof. Totties Gruppe an der BNC-Forschung vertreten.

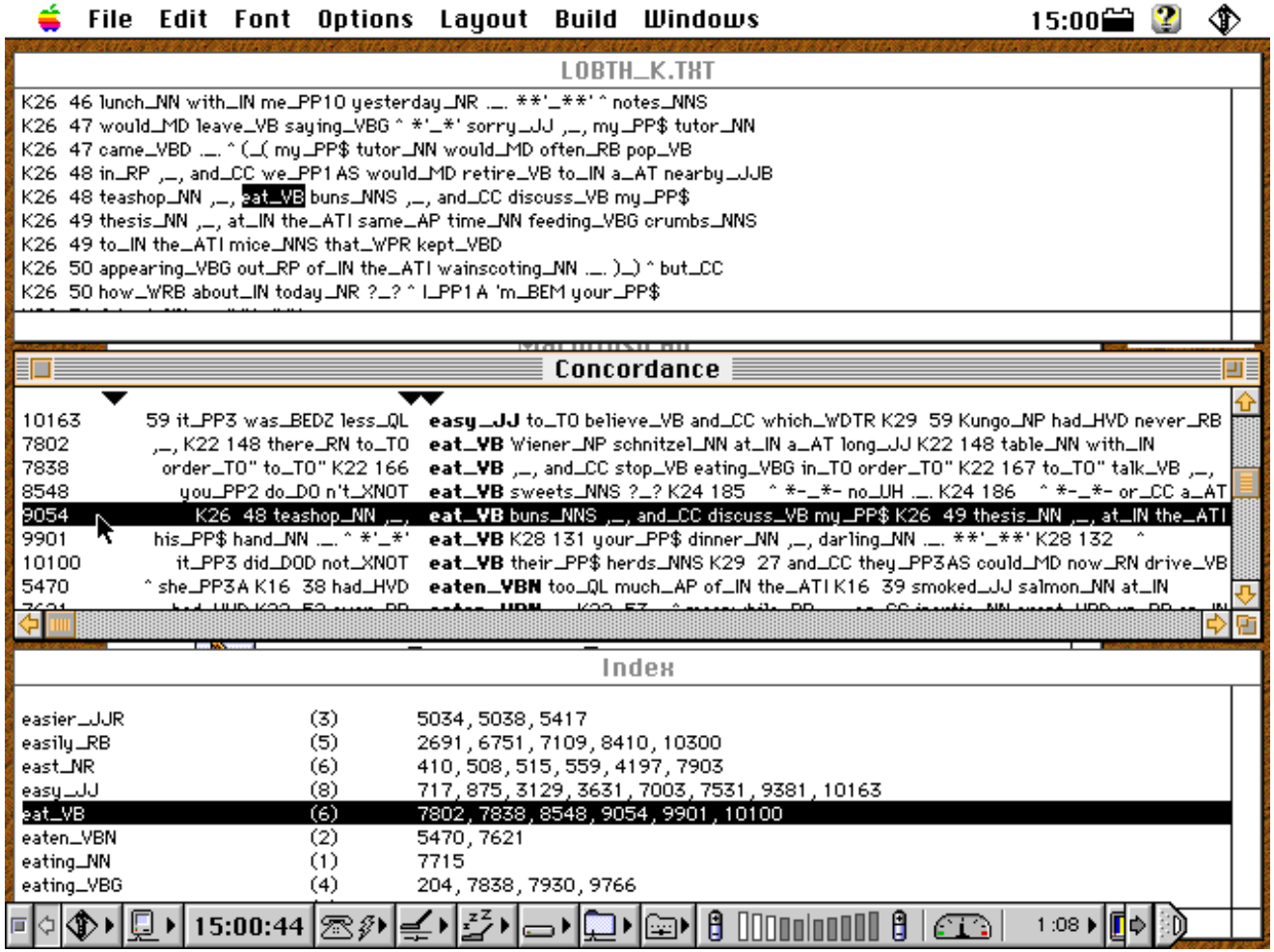
[Kennedy 98] Chapter 2.3 bietet einen Überblick über einige weitere Korpora.

Ein Konkordanzprogramm im Überblick: Conc 1.8 für Macintosh

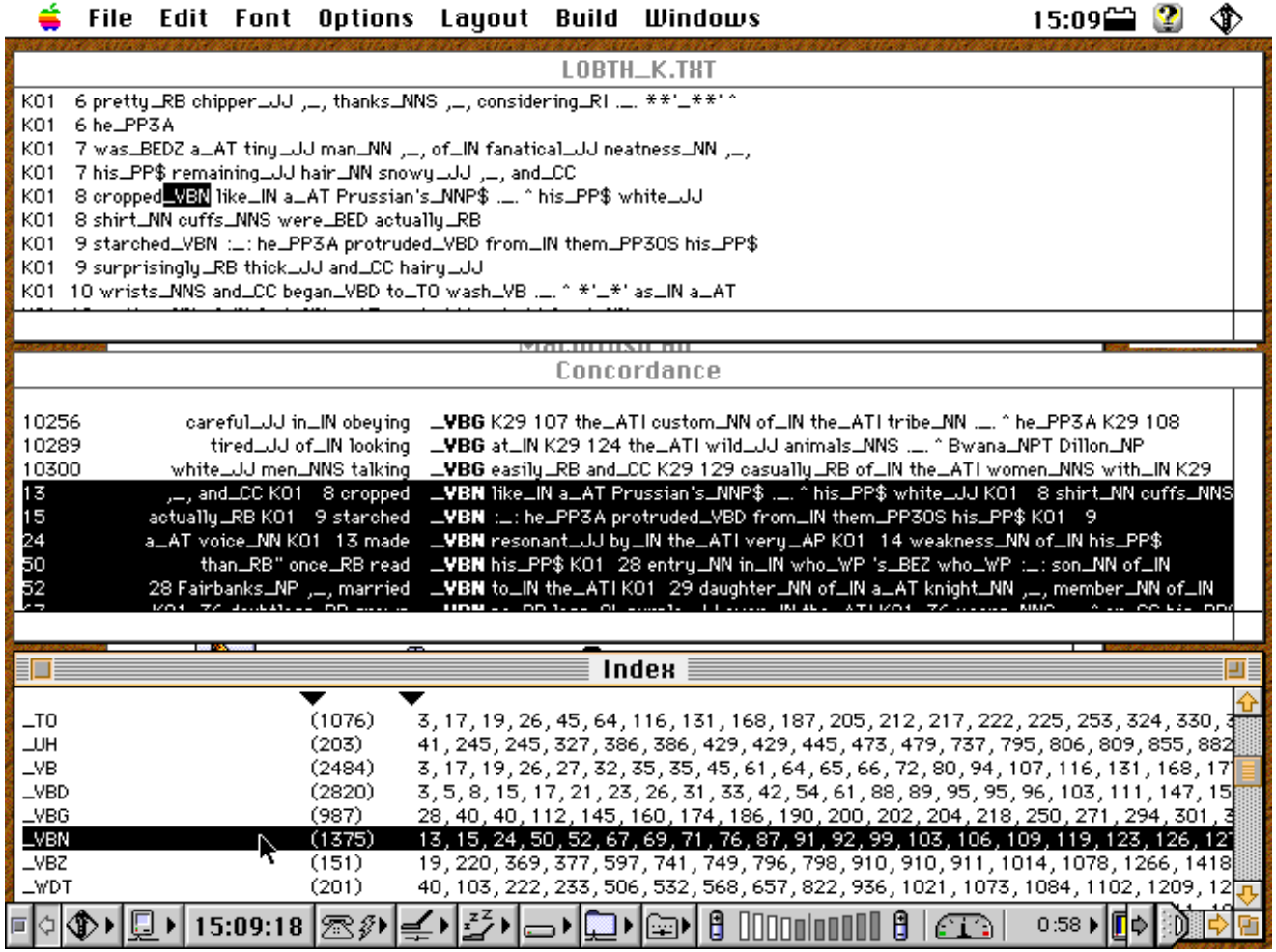
Ein Konkordanzprogramm erlaubt

- das Auffinden aller Verkommnisse eines Suchwortes im Kontext (KWIK)
- das Erstellen von sortierten Wort- und Annotationslisten
- Kollokationsanalysen
- einfache Statistik und tabellarische Ausgabe von Daten für die Weiterverarbeitung

Conc (derzeitge Version ist 1.80b3) kann man frei herunterladen vom amerikanischen Summer Institute of Linguistics (www.sil.org). Es ist relativ einfach, sowohl in der Bedienung als auch in den Möglichkeiten. Es verwendet z.B. Korpora in rohem Textformat, was maximal einfach ist, aber lange Verarbeitungszeiten zum Aufbau der Konkordanz mit sich zieht. Ein Nachteil der Einfachheit ist auch, dass der ganze Korpus im RAM Platz finden muss.



[Conc-Konkordanz mit LOB Korpus Kategorie K, Sortierfolge ASCII]



[Conc-Konkordanz mit LOB Korpus Kategorie K, sortiert nach Tags]

Einfache quantitative Untersuchungen kann man schon in Conc selber vornehmen. So z.B.:

Untersuchungsfrage: Ist wissenschaftlicher Diskurs nominaler als andere Genres?

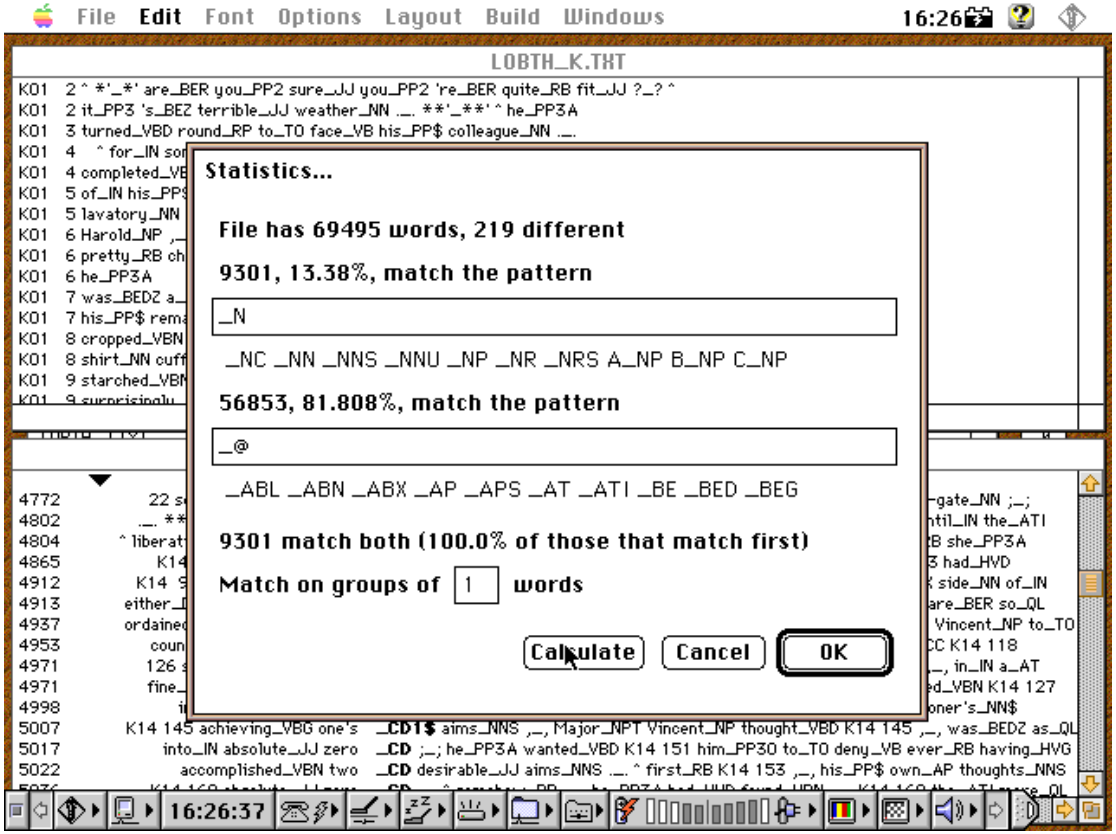
Wir beschränken uns auf einen Direktvergleich der LOB Kategorien K (General Fiction) und J (Learned and Scientific Writing) und versuchen, die folgenden einfachen Fragen zu beantworten:

- Wieviele Prozent aller Worte sind Nomen?
- Wieviele Prozent aller Worte sind Verben?
- Was ist das Verb/Nomen Verhältnis?
- Wie sind die Unterschiede zwischen LOB K und LOB J betreffend dieser Fragen?

→ Zur on-line Demo

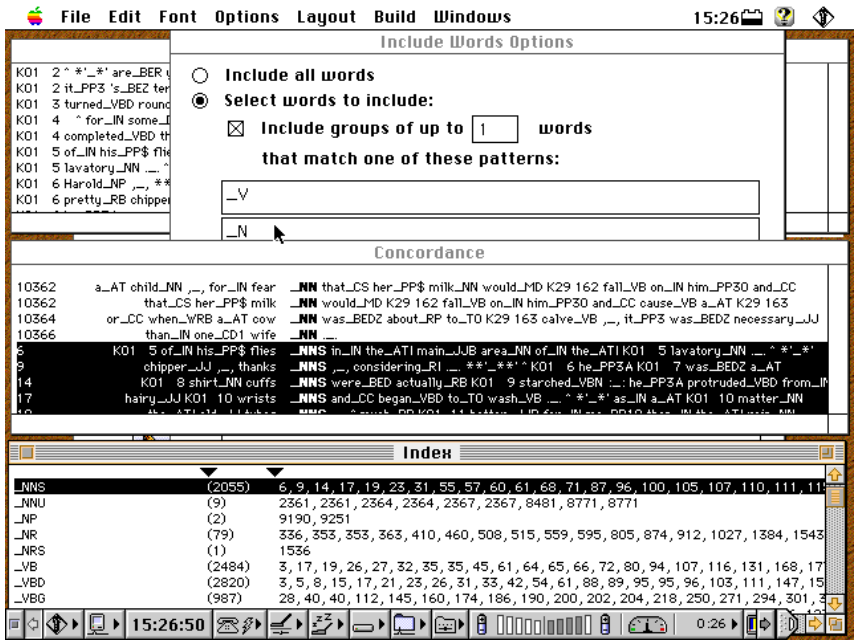
Grob gesagt beginnen im LOB Tagset Verbtags mit **V** und Nomentags mit **N**. Die einfache in Conc eingebaute Statistikabfrage über reguläre Ausdrücke liefert uns dazu folgende Ergebnisse:

- Wieviele Prozent aller Worte sind Nomen?

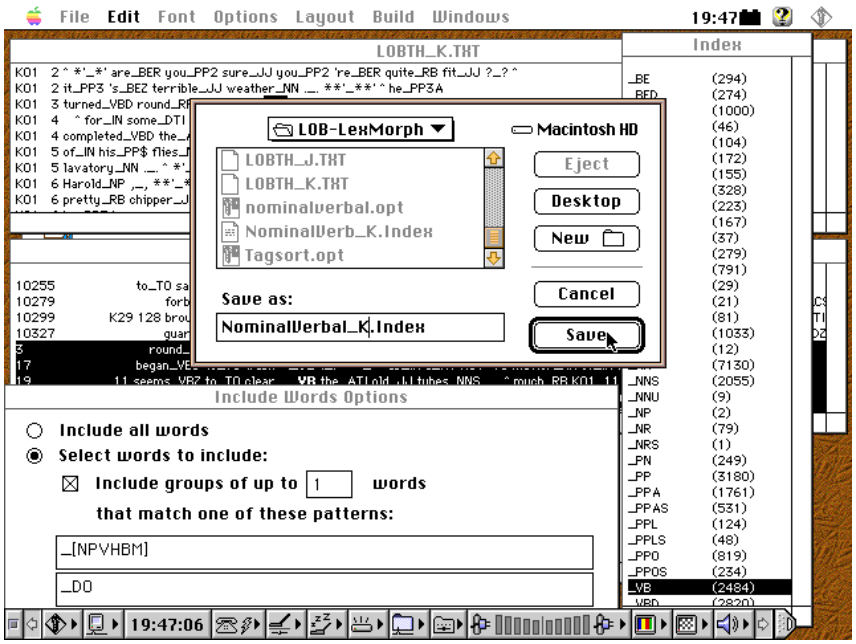


In der online-Demo werden wir

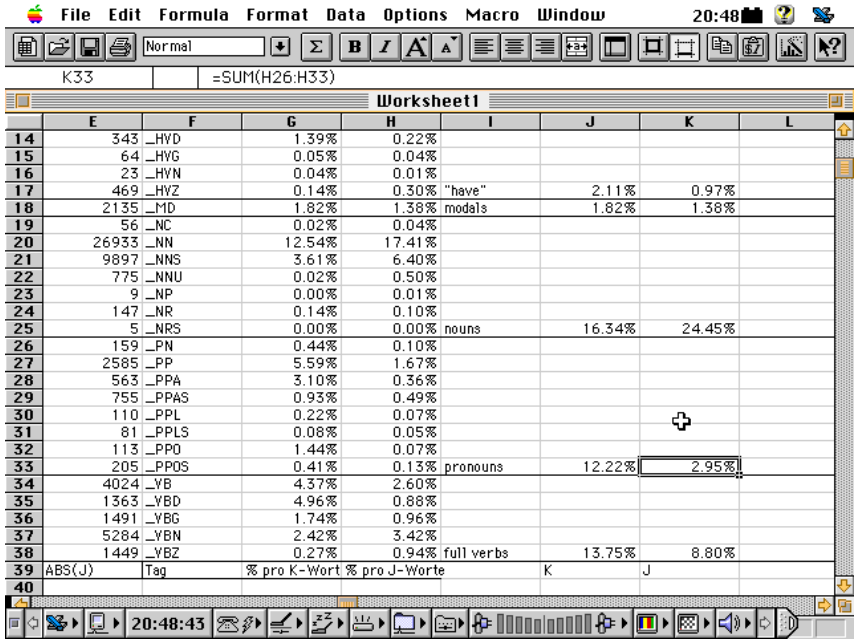
- Einen spezialisierten Index aufbauen



- Diesen als Rohtext exportieren

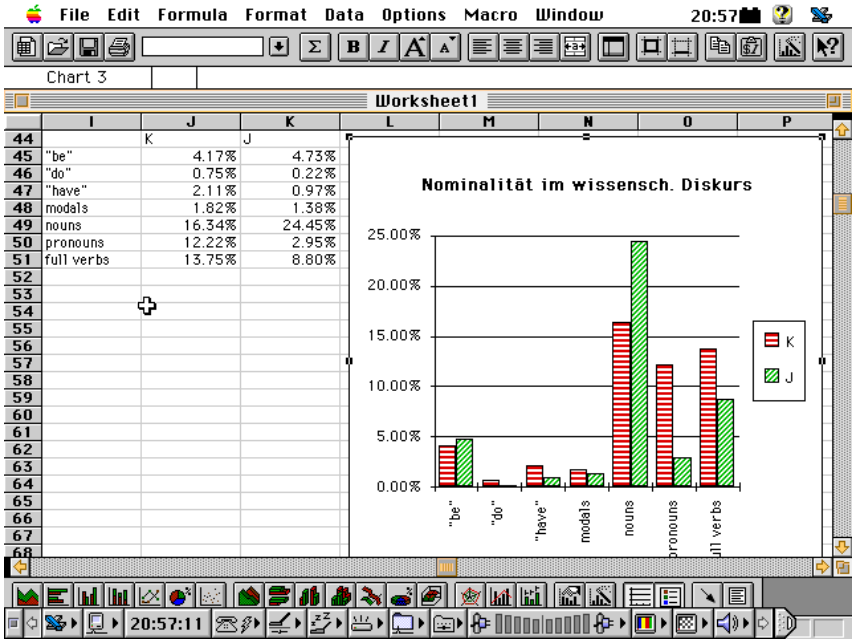


· In Excel tabellarisch verarbeiten



	E	F	G	H	I	J	K	L
14	343	_HVD	1.39%	0.22%				
15	64	_HVG	0.05%	0.04%				
16	23	_HVN	0.04%	0.01%				
17	469	_HVZ	0.14%	0.30%	"have"	2.11%	0.97%	
18	2135	_MD	1.82%	1.38%	modals	1.82%	1.38%	
19	56	_NC	0.02%	0.04%				
20	26933	_NN	12.54%	17.41%				
21	9897	_NNS	3.61%	6.40%				
22	775	_NNU	0.02%	0.50%				
23	9	_NP	0.00%	0.01%				
24	147	_NR	0.14%	0.10%				
25	5	_NRS	0.00%	0.00%	nouns	16.34%	24.45%	
26	159	_PN	0.44%	0.10%				
27	2585	_PP	5.59%	1.67%				
28	563	_PPA	3.10%	0.36%				
29	755	_PPAS	0.93%	0.49%				
30	110	_PPL	0.22%	0.07%				
31	81	_PPLS	0.08%	0.05%				
32	113	_PPO	1.44%	0.07%				
33	205	_PPOS	0.41%	0.13%	pronouns	12.22%	2.95%	
34	4024	_YB	4.37%	2.60%				
35	1363	_YBD	4.96%	0.88%				
36	1491	_YBG	1.74%	0.96%				
37	5284	_YBN	2.42%	3.42%				
38	1449	_YBZ	0.27%	0.94%	full verbs	13.75%	8.80%	
39	ABS(J)	Tag	% pro K-Wort	% pro J-Worte		K	J	
40								

· Als Diagramm ausgeben



	I	J	K
44		K	J
45	"be"	4.17%	4.73%
46	"do"	0.75%	0.22%
47	"have"	2.11%	0.97%
48	modals	1.82%	1.38%
49	nouns	16.34%	24.45%
50	pronouns	12.22%	2.95%
51	full verbs	13.75%	8.80%