CZECH NATIONAL
CORPUS

Introduction to Text Corpora and Their Applications

Corpora in grammar and diachronic studies

Lucie Chlumská, Ph.D.

lucie.chlumska@korpus.cz

OUTLINE:

1. LECTURE

- revision: grammar studies B.C. (before corpus:)
- advantages of corpus-based approach in grammar
- grammatical variation and grammatical change
- lexico-grammar

2. SEMINAR

- reading (Ute Römer): *The Inseparability of Lexis and Grammar*
- patterns in language and how to analyse and describe them

LECTURE

Historical review

The beginnings

First attempts to collect data similar to corpora (before 1960s) were made in the following areas:

- biblical and literary studies
- lexicography
- dialect studies
- language education studies
- grammatical studies

History

- Otto Jespersen, Danish professor, is said to have his villa filled with shoeboxes containing hundreds of thousands of paper slips with examples of interesting English sentences
  - monumental work *A Modern English Grammar on Historical Principles* (1909–49)

- Charles C. Fries used a corpus of letters written to the US government by persons of different educational and social backgrounds to demonstrate social class differences in usage in his *American English Grammar* (1940)
- Later in *The Structure of English* (1952) he used a 250,000-word corpus of recorded telephone conversations
- he analyzed all his corpora manually...

SEU

- the most important pre-electronic corpus was the *Survey of English Usage Corpus* (*SEU*) by Randolph Quirk et al. in 1968
- it marked a transition between earlier non-computerized corpus-based description and the rise of corpus linguistics
- founded in 1959 by Quirk, the SEU aimed to collect 200 samples (each about 5,000 words) representative of both written and spoken language > corpus of 1 million words
  - SEU Corpus contains texts produced between 1953 and 1987, originally available in the form of paper slips filed at the University College London
  - there was a slip for every word in the corpus, containing 17 lines of text plus a mark-up (grammatical features, prosody...)
  - basis for the *A Comprehensive Grammar of the English Language* (1985)

LGSWE

- a new milestone after Quirk's Grammar
- Biber et al. (1999): *Longmann Grammar of Spoken and Written English*
    - a six-year research project, international research team
    - based on the 40-million-word Longman Spoken and Written English Corpus
    - thorough description of English grammar
    - exploring the differences between written and spoken grammar
    - taking register variation into account:
        - conversation
        - fiction
        - news
        - academic prose

CGE

- Carter &McCarthy (2006): *Cambridge Grammar of English*
  - a seven-year research project
  - informed by the 700-million-word Cambridge International Corpus, incl. CANCODE (Cambridge and Nottingham Corpus in Discourse in English, spoken corpus, 5 million)
  - for those interested in ESL (English as a Second Language)
  - many examples from speech to balance out the predominance of written language description in traditional grammars

  - criticised (by e.g. Rodney Huddleston) for "being inconsistent and confusing"

Advantages of corpus-based approach

Benefits of using corpora

- large amount of authentic data, both written and spoken
- grammatical tagging: POS-tagging, syntactic parsing, semantics...
  - very useful, but always an interpretation
  - treebanks – much more difficult nut to crack
  - Gilquin (2002): *corpus grammarians tend to choose topics for research that can be investigated through relatively simple corpus searches, rather than those requiring a high level of abstract structure...*
- software functions:
  - enable to search for patterns or constructions using regular expressions and CQL (= corpus query language), making use of annotation (lemma, tag) at the same time
  - building a complex query can be...complex

Example of a complex CQL query

How to search for an indicative verb form in Czech: by combining the following queries

Future:

```
[word="(?i)(ne)?bud(u|e[šmt]?e?|ou)"]
```

Present:

```
[lemma="být" & word="(?i)((ne)?jsem|(ne)?jseš|(ne)?jsi|je|nen[íi]|(ne)?jsme|(ne)?jste|(ne)?jsou|sem|seš|si|sme|ste|sou)"]
[tag="VB.*" & lemma!="být|_být.*"]
```

Preterite:

```
[lemma="_být" & word="(?i)jsem|jsi|jsme|jste|sem|si|[sz]me|ste"]
[tag="Vp.....2.*" & word="(?i).+l[aoiy]?s"]
[tag="Vp.....3.*"] within (([pos="[JZ]" & lemma!="aby|kdyby"]|<s>) [tag!="Vc.*" & pos!="[ZJ]"]*
[tag="Vp.....3.*"] [tag!="Vc.*" & pos!="[ZJ]"]* ([pos="[JZ]"]|</s>))
```

Contributions of CL to grammar studies

1. unexpected findings

- Hudson (1994): "About 37 % word-tokens are nouns."
  - all varieties of written English show a fixed percentage of nominals
  - "nouns" in this study include pronouns as well (7 % in informative, 14 % in imaginative texts), but combination of pronouns and nouns is a constant

- Tootie and Hoffmann (2006): tag questions nine times more common in BrE conversation than in AmE conversation
  - vast discrepancy, no convincing explanation has yet been found...

Contributions of CL to grammar studies

2. peripheral areas opened up

- principal of total accountability (everything matters and has to be researched)
- adverbials:
  - one of the largest chapter in Quirk et al. (1985) and Biber et al. (1999)
  - distinction between adjuncts (the more central and frequent category of adverbials) and discjuncts and conjuncts (more periplheral)
- discourse markers:
  - also called "pragmatic markers", "discourse particles" etc.
  - uncertain category, straddling the border between grammar, pragmatics and discourse analysis (Aijmer 2002)

Contributions of CL to grammar studies

3. investigating spoken English

- grammar has traditionally focused on the written language (the etymology in class. Greek refers to written symbols and the art of reading and writing)
- spoken language a very hot topic in corpus linguistics with the arrival of new data
  - Tottie (1991): research on negation > roughly twice as common in speech than in writing (due to interactive, involved character of speech)
  - Carter and McCarthy (1995, 2006): grammatical structures unique to speech (initial ellipsis, topics in pre-clause slots, tails in post-clause slots)

"*North and South London – they're different worlds – aren't they in a way?*"

Cambridge Grammar of English

Three main areas of grammatical research

– an increasing amount of research in the last 20 years in the following areas:

1. Grammatical variation

2. Grammatical change

3. Lexico-grammar

Grammatical variation

Grammatical variation

- includes both variationist approach (aimed at variants competing against each other) and text-linguistic research (exploring text frequencies of particular constructions in corpora)

1. variationist sociolinguistics (language internal and external factors)
2. diachronic linguistics
3. register/genre/text type analysis (Biber)
4. dialectology
5. knowledge, processing, cognition (Gries)

Grammatical change

Grammatical change

- change in the regularities that characterize a language system at a given point in time
- comparison of older and more recent stages of language
  - most straightforward grammatical change is when an altogether new option emerges in the system of grammatical forms as the outcome of grammaticalization (*the meat is being cooked* v. *the meat is cooking*)
  - significant shifts in frequencies (*I haven't the time* v. *I don't have the time*).
  - shifts in usage (genres or text types)
- the dividing line between grammaticalization, frequency change and style change is not always sharp
- hallmark of a corpus-based approach is that the grammatical phenomenon is studied in its entirety (all relevant examples are retrieved from a corpus – not just good examples)

Examples of grammatical change

Leech, G. (2003): ‘Modality on the move: the English modal auxilieries 1961-1992‘, In *Modality in Contemporary English*, Berlin: Mouton de Gruyter, 223-240.

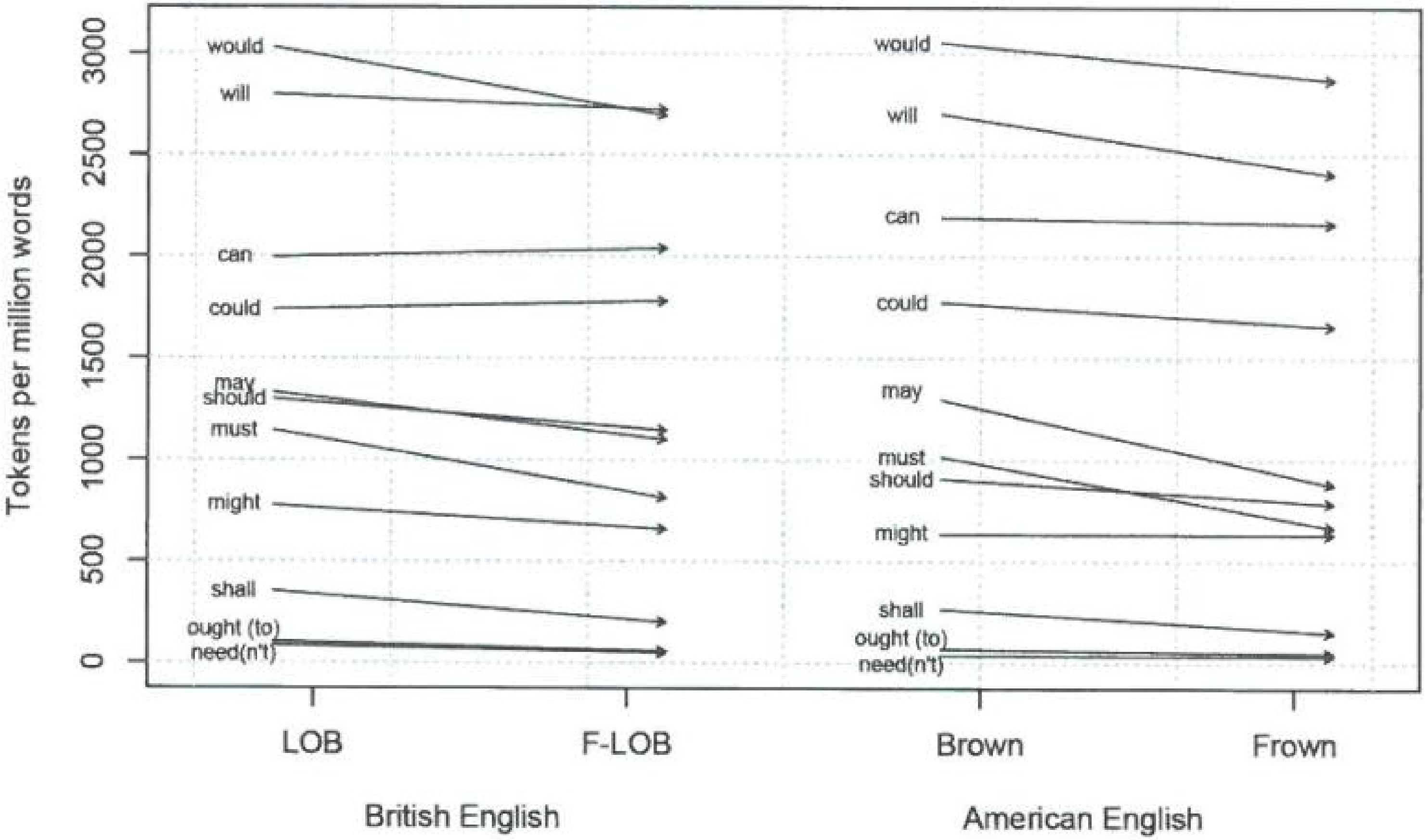
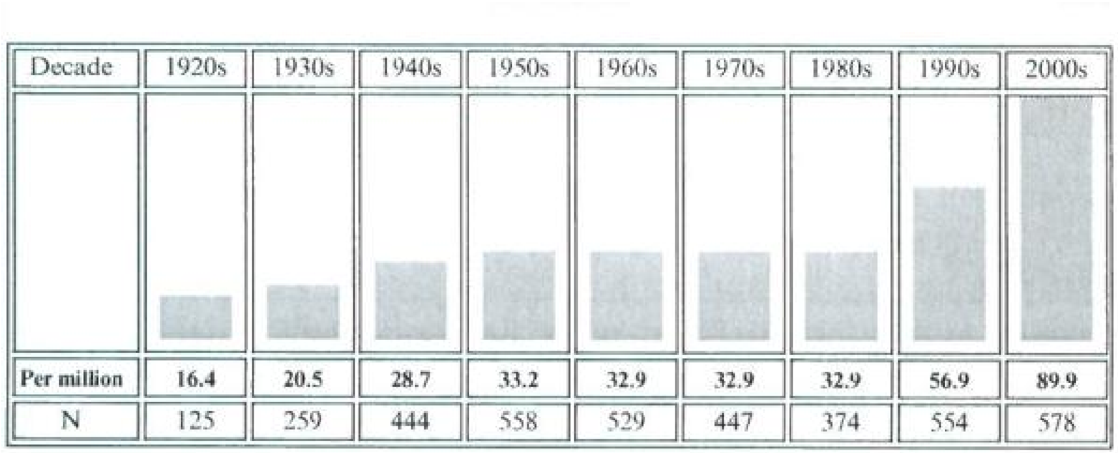


Figure 10.2 The decline of the modal auxiliaries (based on Leech 2003: 228, table 3)

Examples of grammatical change

Mair, C. (2006): ‘ Tracking ongoing grammatical change and recent diversification in present-day standard English: The complementary role of small and large corpora‘, in *The Changing Face of Corpus Linguistics*, Amsterdam: Rodopi, 355-376.



Decade	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
Per million	**16.4**	**20.5**	**28.7**	**33.2**	**32.9**	**32.9**	**32.9**	**56.9**	**89.9**
N	125	259	444	558	529	447	374	554	578

Figure 7.3 Frequency of *get*-passives in the Time Corpus: per million words.

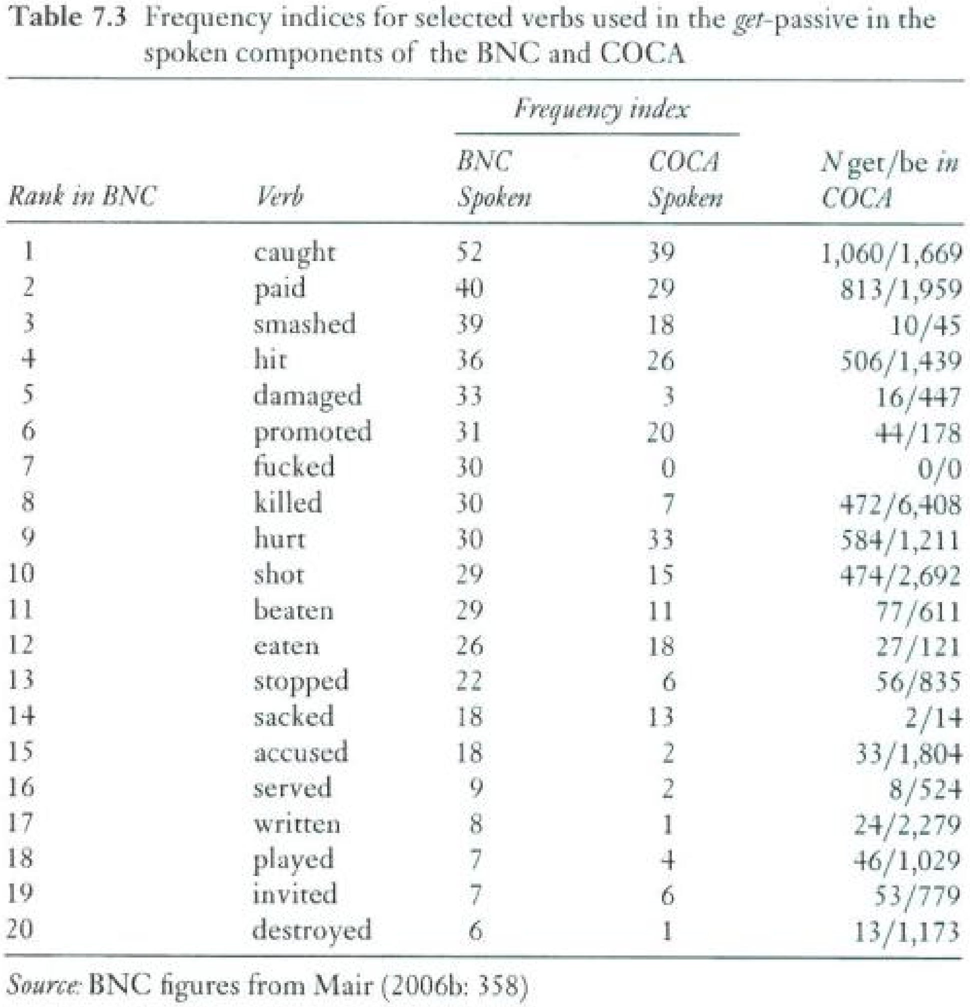


Table 7.3 Frequency indices for selected verbs used in the *get*-passive in the spoken components of the BNC and COCA

		Frequency index		
Rank in BNC	*Verb*	*BNC Spoken*	*COCA Spoken*	*N* get/be *in COCA*
1	caught	52	39	1,060/1,669
2	paid	40	29	813/1,959
3	smashed	39	18	10/45
4	hit	36	26	506/1,439
5	damaged	33	3	16/447
6	promoted	31	20	44/178
7	fucked	30	0	0/0
8	killed	30	7	472/6,408
9	hurt	30	33	584/1,211
10	shot	29	15	474/2,692
11	beaten	29	11	77/611
12	eaten	26	18	27/121
13	stopped	22	6	56/835
14	sacked	18	13	2/14
15	accused	18	2	33/1,804
16	served	9	2	8/524
17	written	8	1	24/2,279
18	played	7	4	46/1,029
19	invited	7	6	53/779
20	destroyed	6	1	13/1,173

Source: BNC figures from Mair (2006b: 358)

Lexico-grammar

Research context

- grammar-to-lexis viewpoint

– takes grammar categories as prior and notes the lexis that occurs disproportionately frequently in each category

  – *Longman grammar of spoken and written English*

  – collostructions (constructions + collocations) (Stefanowitsch & Gries)

- lexis-to-grammar viewpoint

– takes lexis as prior and notes the frequently occurring grammatical contexts of each word

  – Pattern Grammar (Sinclair, Francis, Hunston)

  – Lexical priming (Hoey)

Thank you for your attention!

Questions?

Let's talk about language!

CZECH NATIONAL
CORPUS

SEMINAR

Reading

common reading:

Römer, U. (2009). “The Inseparability of Lexis and Grammar. Corpus linguistic perspectives.” In *Annual Review of Cognitive Linguistics*, Volume 7, pp 140–62.

Discussion

- What does a corpus-based approach to grammar focus on?
- What is an idiom principle?
- What is a pattern in language and how can it be retrieved from a corpus?
- What is a lexical bundle and how does it differ from a collocation?
- Why to look at grammar and lexis at the same time?
- Are there any limitations of such an approach?