Corpora in Lexicography

Overview
1. An experiment
2. Lecture on "Lexikologie, Lexikographie und Lexikonstrukturen":
   http://www.ifi.unizh.ch/CL/volk/LexMorphVorL/Lexikon09.Lex.html
3. Special projects

How would you make a dictionary entry?
- for the token 'einstellen'
- for the token 'Natel'

Computer use for lexicography
- display context (concordances, KWIC)
- database for lexical entries

Computer use for lexicography
- count tokens in corpus
- count neighboring tokens (collocations)
- find multiword units
  - Idioms
  - names
  - discontinuous elements
- determine thesaurus relations

- http://www.dwds.de/ (Berlin-Brandenburgische Akademie der Wissenschaften)
- Kollokationen: Dies sind Wortverbindungen, die in gewisser Weise eine lexikalische Einheit bilden, beispielsweise
  - Aufmerksamkeit zollen,
  - zur Entscheidung bringen,
  - in die Quere kommen,
  - mit Kind und Kegel,
  - freie Marktwirtschaft,
  - schwach tendieren,
  - gut und gerne.
The "Wortwarte" project

- at the University of Tübingen (Lothar Lemnitzer, Tylman Ule)
- http://www.sfs.nphil.uni-tuebingen.de/~lothar/nw/index.html
- a monitor corpus project for German
- daily download of newspapers (Die Zeit, die Welt, Financial Times Deutschland, Rheinische Post etc.).
- comparison of the words contained in these papers against a word list of the "Deutsches Referenzkorpus", a corpus that consists of 120 million tokens corresponding to 2.3 million types.

What is found:
- misspelled words
- words with unusual spelling (e.g. Betel-Nuss instead of Betelnuss) (partly because of the Rechtschreibreform)
- regular words that are not in the reference corpus by accident
- words from the spoken language with no fixed spelling (boahh, iiiih etc.)
- new words

New words are
- seldom really new. Sometimes based on product names ("Nogger Dir einen")
- Derivation ("faxen")
- Compounding (most often: "überfaxen")
- loan words (mostly from English)
- abbreviations

Reasons for new words
- new things are being introduced (Handy)
- new words sound better than their predecessors (Banker vs. Bankangestellter)

Neologisms of Dec. 9, 2002
- Erdabplattung
- Internetregister
- Keksdesigner
- Over-the-Counter-Geschäft
- Petaflop
- zielgruppenaffin

Grammaticography
- find verbal subcat frames
  - accusative, dative, genitive, prepositional objects but also sentential complements (zu-infinitive, dass-clause, wh-clause)
- with their respective frequencies
German verbal subcat frames

Approach using a corpus
- PoS tagging
- Chunk parsing
  - with case assignment
  - with active/passive assignment

Example:
- weil der Manager der Chemiefirma xxxt, dass die Fusion in zwei Monaten stattfindet.
- weil [der Manager nom] [der Chemiefirma gen/dat] xxxt, dass die Fusion in zwei Monaten stattfindet.
Possible frames:
  - subj + s-comp(dass)
  - subj + sobj(dative) + s-comp(dass)

Approach
- Use only unambiguous corpus examples
  - e.g. with clear case marking for genitives
  - without nouns that subcategorize dass-clauses
- find nouns that subcategorize dass-clauses by checking them in sentence-initial position (V-second clauses).
  - requires very large corpora

Approach
- Nominative must be clearly marked (via determiners or pronouns)
- Accusative/nominative is then resolved as accusative. Such an NP must not contain a temporal head (Er liest den ganzen Tag.)
- Genitive NPs must not contain a temporal head (Er kam eines Tages vorbei.)

Result
- 6305 Verb lemmas
- 244 different subcat frames (on average 1.9 frames per verb lemma)
- comparison against DUDEN Gesamtwörterbuch for 15 verbs from different frequency ranks
  - the system finds a lot more sentential subcat frames
  - the dictionary lists some frames that did not appear in the corpus