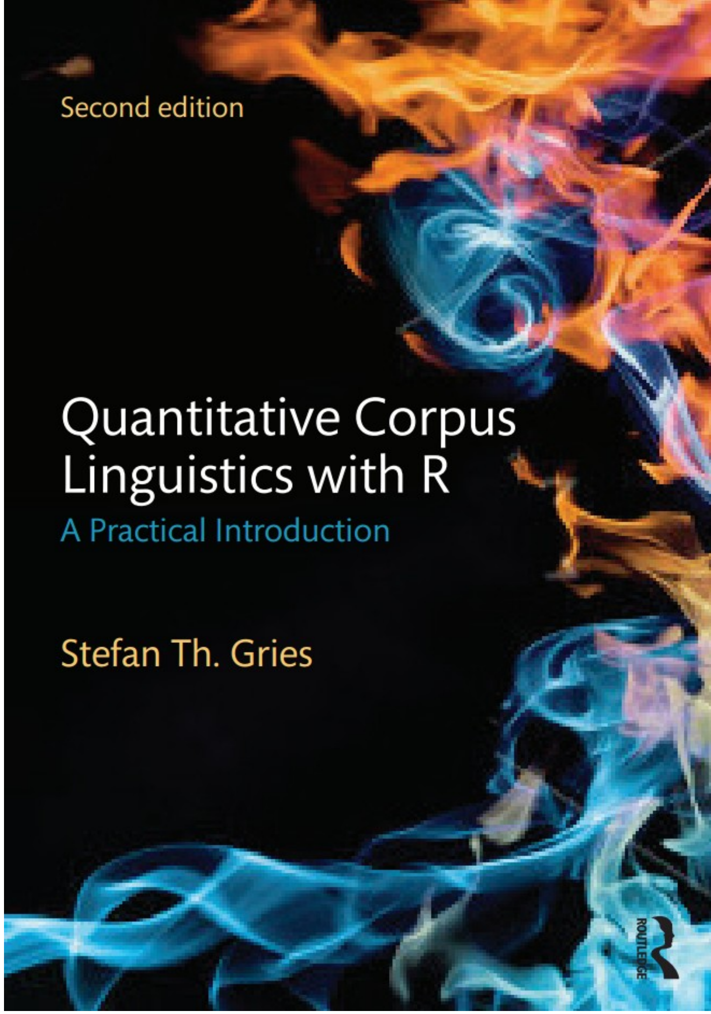
Corpus Linguistics

Stefan Th. Gries
UC Santa Barbara &
JLU Giessen



Other kinds of data used in linguistics

- Corpora with written texts (newspapers, blogs, …)
- text archives, example collections
- corpora with recorded spoken language in communities where recording is not extraordinary
- corpora with recorded spoken language in communities where recording is extraordinary
- interview data
- experimentation where subjects do something with language that they usually do anyway
  - answering questions in priming studies
  - picture descriptions in studies on information structure
- elicited data from fieldwork ("how do you say X in your language?")
- experimentation where subjects do something with language that they usually do not anyway
  - involving units they typically interact with
    - sentence acceptability judgments, sentence sorting, lexical decision tasks
  - involving units they typically do not interact with and linguistic output
    - phoneme monitoring, gating
  - involving units they typically do not interact with and non-linguistic output
    - event-related potentials, eye-movements, ultrasound tongue-position videos

What is a corpus?

- A prototypical corpus is
  - a machine-readable collection of (spoken or written) texts
  - that were produced in a natural communicative setting
  - representative and balanced with respect to a particular variety/register/genre
  - that are compiled with the intention to be analyzed linguistically

What is a corpus?

- "Machine-readable"
  - virtually all corpora are stored in the form of plain text files (ASCII or Unicode) that can be loaded, manipulated, and processed platform-independently
  - frequent formats of annotated corpora
    - SGML
    - XML
    - what you will not find much (anymore)
      - corpus files as *.doc
      - corpus data on paper
  - some corpora come with sophisticated retrieval software (e.g., ICE-GB)

What is a corpus?

- "produced in a natural communicative setting"
  - the texts were spoken or written for some authentic communicative purpose, not for putting them into a corpus
  - example: journalese in corpora meets this criterion
    - journalists write articles to communicate something in their newspapers, not to fill a linguist's corpus
  - example: if I record someone's speech for a week, I will hopefully obtain authentic discourse (even though all interlocutors should know they are being recorded)

What is a corpus?

- "representative with respect to a particular variety"
  - the different parts of the variety I am interested in are all manifested in the corpus
  - example: phonological reduction in the speech of Californian adolescents
    - if I only record Californian adolescents in their peer groups, I would fail to collect data on a whole variety of additional groups of interlocutors
      - parents
      - teachers
      - …

What is a corpus?

- "balanced with respect to a particular variety"
  - not only should all parts of the variety I am interested in be included
  - also, the proportions of the parts with which they are represented in the sample (i.e., the corpus) should reflect the proportions with which they occur in the population
  - example: if dialogs make up 65% of the speech of Californian adolescents, 65% of my corpus of the speech of Californian adolescents should be dialog data

What is a corpus?

- "balanced with respect to a particular variety"
  - problems
    - we can only measure a small sample of Californian adolescents: the percentage will vary
    - how would we measure the proportions: in terms of minutes, words, sentences, …?
    - how would we measure the importance of any linguistic variety?
      - usually, conversational speech is considered primary …
      - … but a single newspaper headline may have a more radical effect on any speaker's linguistic system than many hours of conversational speech
  - balancedness = theoretical ideal

What is a marginal corpus?

- A collection of second/foreign language learner essays
  - such essays are usually not produced in a natural communicative setting
    - teachers assign topics
    - teachers impose time limits
    - teachers impose word limits
    - teachers grade

What is not a corpus?

- A text archive: a database of texts
  - which were often not produced in a natural setting
  - which was usually not compiled for linguistic analysis
  - which was not intended to be representative of any particular linguistic variety
  - which was not intended to be balanced with respect to any particular linguistic variety
  - example: a publisher of some popular computing periodical makes all issues of the previous year available on a website

What is not a corpus?

- An example collection of words, sentences, …
  - example: I am currently compiling a collection of morphological blends (*brunch*, *motel*, …)
  - example: psycholinguists have collected huge numbers of speech errors
  - such an example collection is
    - probably not balanced since some errors are difficult to notice in the first place
    - probably not representative since each individual's range of interlocutors is very small and does not cover the whole range of possible interlocutors

What kinds of corpora are there?

- Some central distinctions
  - general vs. specific
  - static vs. dynamic
  - diachronic vs. synchronic
  - raw vs. annotated
  - monolingual vs. parallel

What kinds of corpora are there?

- General vs. specific
  - general corpora aim at being representative for 'a language as a whole'
    - British National Corpus
    - ANC
    - RNC
    - …
  - specific corpora aim at being representative for a particular variety of a language
    - language acquisition corpora in the CHILDES database
    - COLT: The Bergen Corpus of London Teenager Language
    - The Wall Street Journal Corpus
    - …

What kinds of corpora are there?

- Static vs. dynamic
  - static corpora have a fixed size: once the corpus compilation up to that size is completed, the corpus remains as it is (in terms of the samples it contains)
  - dynamic corpora are continuously changing as new material is added to them all the time
  - the issue is 'quality vs. quantity'

What kinds of corpora are there?

- Diachronic vs. synchronic
  - diachronic corpora aim at providing data on a longer stretch of development of a (variety of a) language over time
  - synchronic corpora aim at providing a snapshop of a language at point X (where a "point" may well comprise a decade)

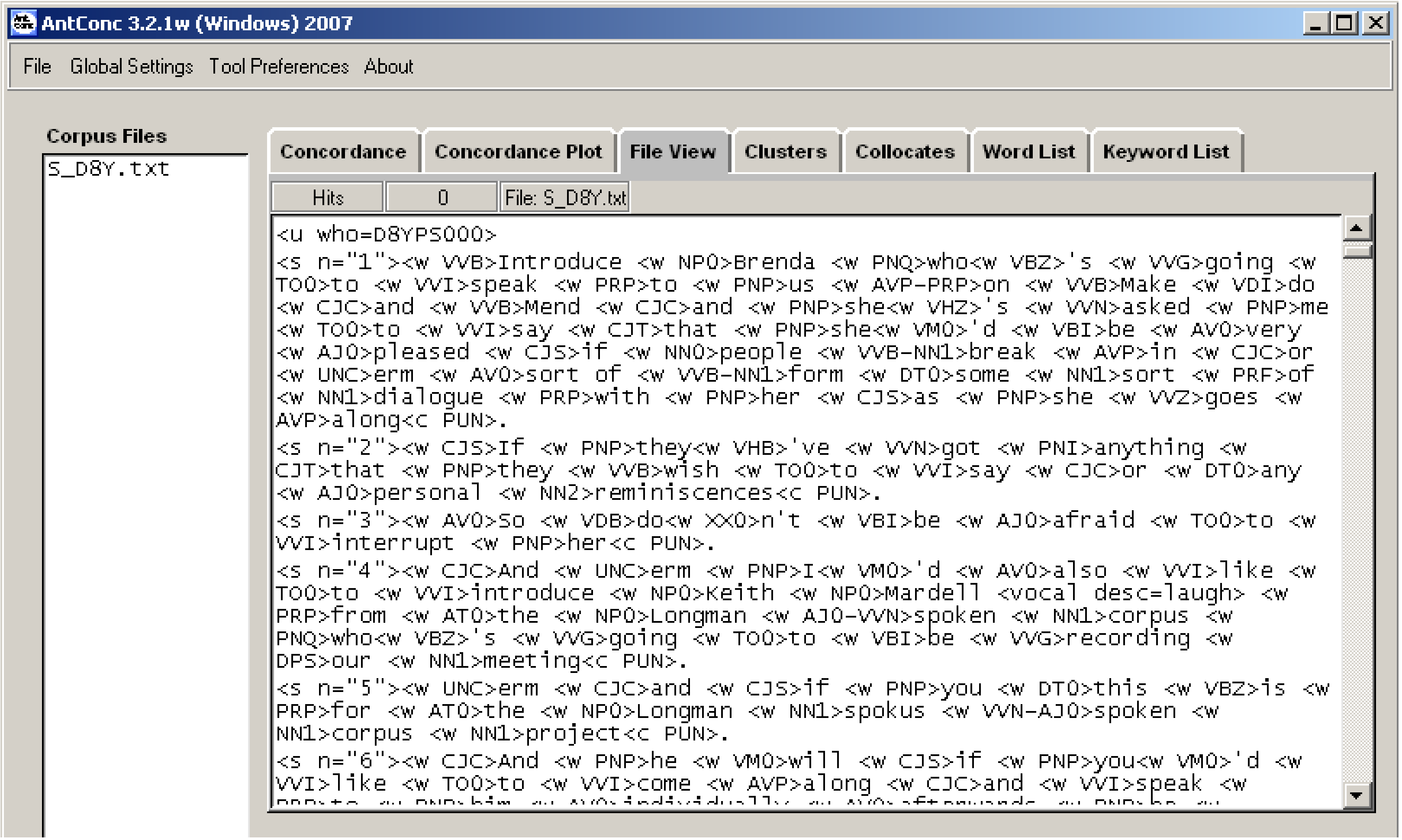
What kinds of corpora are there?

- Raw vs. annotated
  - raw corpora contain only the actually produced texts/utterances (usually with some sort of markup such as sources, speaker identifiers, etc.)
  - annotated corpora contain additional information of various kinds; usually, this is information resulting from linguistic or other analysis
    - lemmatization
    - part-of-speech tagging
    - phonological annotation
    - syntactic parse trees
    - multi-level annotation with tiers

What kinds of corpora are there?

- Raw
  - I did get a postcard from him
- lemmatization
  - I<I> did<do> get<get> a<a> postcard<postcard> from<from> him<he>.<punct>
- phonological annotation
  - [@:] I ^did get a !p\ostcard fr/om him# - -
- POS-tagging
  - I<PersPron> did<VerbPast> get<VerbInf> a<Det> postcard<NounSing> from<Prep> him<PersPron>.<punct>

What kinds of corpora are there?



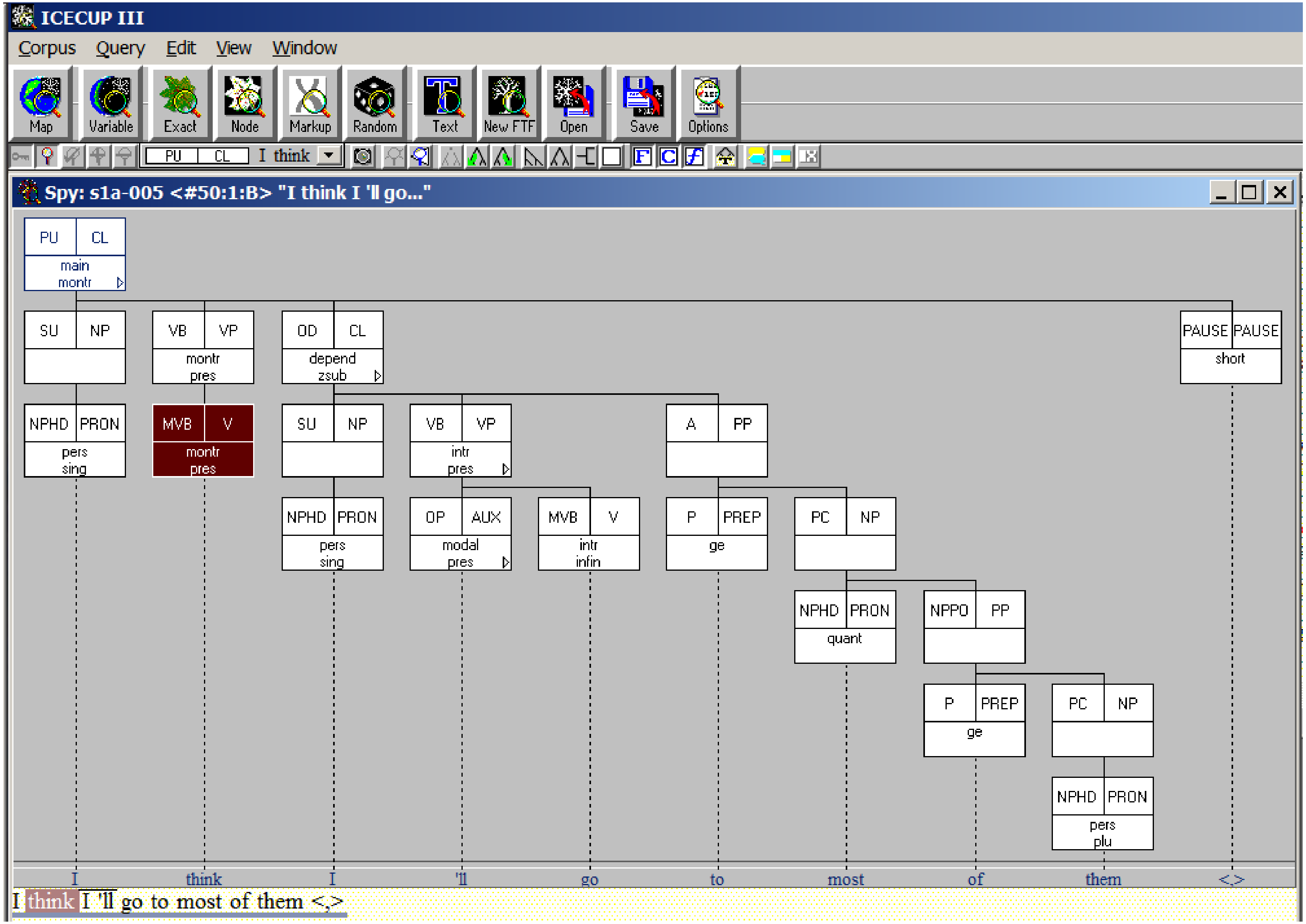
```
<u who=D8YPS000>
<s n="1"><w VVB>Introduce <w NP0>Brenda <w PNQ>who<w VBZ>'s <w VVG>going <w
TO0>to <w VVI>speak <w PRP>to <w PNP>us <w AVP-PRP>on <w VVB>Make <w VDI>do
<w CJC>and <w VVB>Mend <w CJC>and <w PNP>she<w VHZ>'s <w VVN>asked <w PNP>me
<w TO0>to <w VVI>say <w CJT>that <w PNP>she<w VM0>'d <w VBI>be <w AV0>very
<w AJ0>pleased <w CJS>if <w NN0>people <w VVB-NN1>break <w AVP>in <w CJC>or
<w UNC>erm <w AV0>sort of <w VVB-NN1>form <w DT0>some <w NN1>sort <w PRF>of
<w NN1>dialogue <w PRP>with <w PNP>her <w CJS>as <w PNP>she <w VVZ>goes <w
AVP>along<c PUN>.
<s n="2"><w CJS>If <w PNP>they<w VHB>'ve <w VVN>got <w PNI>anything <w
CJT>that <w PNP>they <w VVB>wish <w TO0>to <w VVI>say <w CJC>or <w DT0>any
<w AJ0>personal <w NN2>reminiscences<c PUN>.
<s n="3"><w AV0>So <w VDB>do<w XX0>n't <w VBI>be <w AJ0>afraid <w TO0>to <w
VVI>interrupt <w PNP>her<c PUN>.
<s n="4"><w CJC>And <w UNC>erm <w PNP>I<w VM0>'d <w AV0>also <w VVI>like <w
TO0>to <w VVI>introduce <w NP0>Keith <w NP0>Mardell <vocal desc=laugh> <w
PRP>from <w AT0>the <w NP0>Longman <w AJ0-VVN>spoken <w NN1>corpus <w
PNQ>who<w VBZ>'s <w VVG>going <w TO0>to <w VBI>be <w VVG>recording <w
DPS>our <w NN1>meeting<c PUN>.
<s n="5"><w UNC>erm <w CJC>and <w CJS>if <w PNP>you <w DT0>this <w VBZ>is <w
PRP>for <w AT0>the <w NP0>Longman <w NN1>spokus <w VVN-AJ0>spoken <w
NN1>corpus <w NN1>project<c PUN>.
<s n="6"><w CJC>And <w PNP>he <w VM0>will <w CJS>if <w PNP>you<w VM0>'d <w
VVI>like <w TO0>to <w VVI>come <w AVP>along <w CJC>and <w VVI>speak <w
PRP>to <w PNP>him <w AV0>individually <w AV0>afterwards <w PNP>he <w
```

What kinds of corpora are there?

- Raw
  - I did get a postcard from him
- syntactic parse trees

```
<Subject, NP>
  I<PersPron>
<Predicate, VP>
  did<Verb>
  get<Verb>
<DirObject, NP>
  a<Det>
  postcard<NounSing>
<Adverbial, PP>
  from<Prep>
  him<PersPron>
```

What kinds of corpora are there?



I think I 'll go to most of them <,>

What kinds of corpora are there?

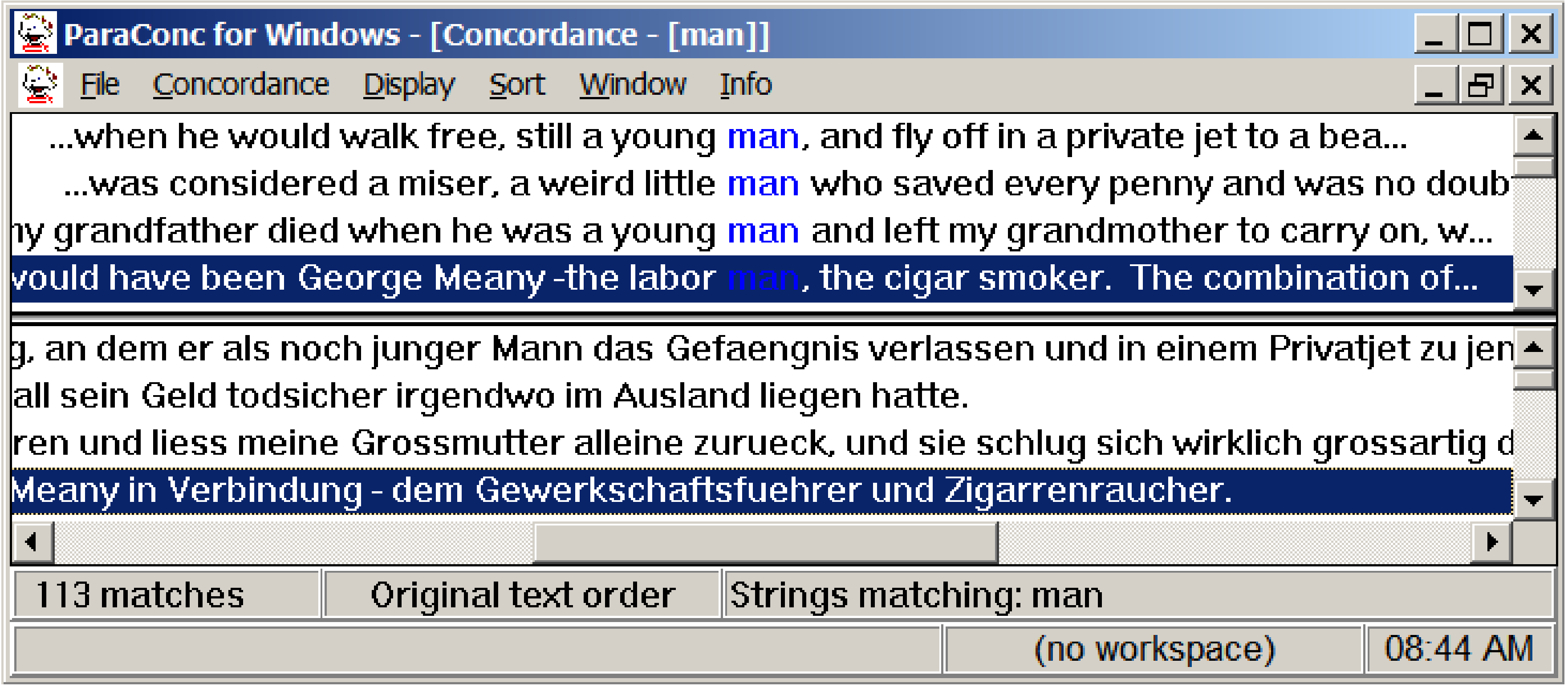
- Raw
  - I did get a postcard from him
- multi-level chat-style annotation

```
*CHI:  I did get a postcard from him
%mor:  pro|I v|do&PAST v|get det|a n|postcard prep|from
       pro|him .
%lex:  get
%syn:  trans
```

What kinds of corpora are there?

- monolingual vs. parallel
  - monolingual corpora are corpora compiled with the intention to provide data for only one language (exceptional data: utterances in a foreign language)
  - parallel corpora provide, ideally, the same text in different languages
    - examples
      - translations from EU Parliament debates
      - Canadian Hansard Corpus (English and French)
    - important issue: alignment of sentences

What kinds of corpora are there?



...when he would walk free, still a young man, and fly off in a private jet to a bea...
...was considered a miser, a weird little man who saved every penny and was no doub
ny grandfather died when he was a young man and left my grandmother to carry on, w...
vould have been George Meany -the labor man, the cigar smoker. The combination of...

g, an dem er als noch junger Mann das Gefaengnis verlassen und in einem Privatjet zu jen
all sein Geld todsicher irgendwo im Ausland liegen hatte.
ren und liess meine Grossmutter alleine zurueck, und sie schlug sich wirklich grossartig d
Meany in Verbindung - dem Gewerkschaftsfuehrer und Zigarrenraucher.

113 matches | Original text order | Strings matching: man

(no workspace) | 08:44 AM

The web as a corpus

- Some advantages
  - new content is constantly added
    - large amounts of data are available
    - linguistic change can be monitored
  - inherently machine-readable
  - universally available
  - freely available
  - diverse data
    - linguistically (many languages)
    - topically (many topics, registers, genres)
    - demographically

The web as a corpus

- Some disadvantages
  - correctness of usage
    - lack of editing
    - native vs. non-native speakers/writers
  - counts
    - pages vs. words
    - multiple copies of identical documents
    - caching of search engines distorts counts
    - non-permanence rules out replicability
  - limited searchability
    - pages vs. words (KWIC)
    - no linguistic annotation
  - representativity and balancedness
    - really demographically diverse? what about underdeveloped or politically restrictive countries? very old people?
    - really topically diverse? isn't the web just about tech, porn, and advertising?
    - prominence of patterns particular to only the internet genre

Kinds of corpora

- Examples from the good ol' pre-computer days
  - Kaeding (1897)
    - 11m words of German to investigate sequences of letters
  - Thorndike (1921)
    - 4.5m words to generate word frequency lists for language learning and teaching purposes
  - Fries (1952)
    - 250,000 words corpus of recorded telephone conversations to develop a grammar of English

Kinds of corpora

- The early days of computers (with punch cards etc.)
  - Brown corpus (Kučera and Francis 1967, rev. ed. 1979)
    - size: 1m words
    - selection of random samples of written AmE
      - 374 samples of 2000+ words of informative prose: press reportage, press editorial, press reviews, religion, skills & hobbies, popular lore, belles lettres, miscellaneous, learned
      - 126 samples of 2000+ words of imaginative prose: general fiction, mystery & detective, science, adventure & western, romance & love, humor
    - several versions differing in annotation

Kinds of corpora

- The early days of computers (with punch cards etc.)
  - Lancaster-Oslo-Bergen corpus (Johansson, Hauge, and Leech 1978)
    - size: 1m words
    - selection of random samples of written BrE
      - 374 samples of 2000+ words of informative prose: press reportage, press editorial, press reviews, religion, skills & hobbies, popular lore, belles lettres, miscellaneous, learned
      - 126 samples of 2000+ words of imaginative prose: general fiction, mystery & detective, science, adventure, romance & love, humor
    - several versions differing in annotation

Kinds of corpora

- Contemporary widely distributed static corpora
  - British National Corpus World edition
    - size: ≈100m words
    - selection of random samples of BrE
      - written texts: 75% informative writing + 25% imaginative writing
      - spoken texts: demographic component containing transcriptions of spontaneous natural conversations (124 people) + context-governed component of recordings made at specific events
  - American National Corpus (100m words, modeled after the BNC, but it's apparently never getting finished)
  - Russian National Corpus (now 65m words, aiming at 100m words)
  - Czech National Corpus (100m words)
  - BYU corpora: Corpus of Contemporary American English, Corpus del Espanol, …

Kinds of corpora

- Generalized monitor corpora
  - The Bank of English (1991): >500m words
  - Corpus of Contemporary American English: >400m words
- more specialized giga corpora
  - The Wall Street Journal Corpus
  - North American News Corpus, News on the Web corpus, …

Compiling corpora

- The central issues
  - size
  - diversity
  - sampling strategy
  - legal stuff
  - annotation

Compiling corpora

- The central issues
  - size
    - object of study: grammatical phenomena vs. lexical phenomena
    - approach: synchronic vs. diachronic
    - data needed: written, spoken, both, more specific register distinctions?
    - practical matters
      - number of samples
      - sampling sizes
      - amount of transcription involved
      - amount of annotation involved
      - other processing issues (e.g., parallel alignment)

Compiling corpora

- The central issues
  - size: questions
    - how long should text samples be to reliably represent the distribution of linguistic features?
      - for many grammatical features, 1,000 words samples appear to be sufficient
    - how many texts in each (e.g., register) category are required to reliably represent the category?
      - for many grammatical features, 10 text samples appear to be sufficient
    - how many texts are needed altogether to accurately identify the salient parameters of variation?
      - for many grammatical features, corpora with 120 texts appear to be sufficient

Compiling corpora

- The central issues
  - diversity
    - situationally defined text categories: registers / genres (e.g., fiction, sports broadcasts, …)
    - linguistically defined text categories: text types (with shared linguistic co-occurrence patterns)
    - dialectal parameters
      - geographic region, age, sex, social class, ethnic groups, education, occupation, …
  - situationally defined $>_{\text{more important}}$ linguistically defined

Compiling corpora

- The central issues
  - diversity: a proposal by Biber (1993)
    - written: published
    - written: unpublished
    - speech: institutional, public, private
    - scripted speech: institutional, public media, other
  - the ICE-GB is largely based on this sampling scheme

Compiling corpora

- The central issues
  - sampling sizes
  - sampling strategy
    - proportional sampling (to achieve balancedness)
    - stratified sampling (to achieve broad coverage)
    - phenomenon-dependent (cf. above)

Compiling corpora

- The central issues
  - legal stuff
    - copyright issues
    - permissions to do recordings from recorders and their interlocutors

```
<s n="4">  And erm I'd also like to introduce Keith Mardell
           <vocal desc=laugh> from the Longman spoken corpus
           who's going to be recording our meeting.
<s n="5">  erm and if you this is for the Longman spoken corpus
           project.
<s n="6">  And he will if you'd like to come along and speak to
           him individually afterwards he will tell you
           something about that.
```

(BNC World edition: D8Y)

    - anonymization: names of people and locations, dates, …

Compiling corpora

- The central issues
  - markup and annotation
    - which markup and annotation is needed? (cf. above)
    - how comprehensive should it be?
    - how can or should it be done?
      - automatically
      - semi-automatically
      - manually
    - for parallel corpora
      - annotation: cf. above
      - how should the alignment be done?

Why use corpora in the first place?

- Chomsky's criticism
  - linguists study competence, not performance
  - language is an unlimited system, corpora are finite
  - 95% of all utterances are ungrammatical anyway
  - corpora are unnecessary and you need judgment data anyway
  - corpus analysis are too time-consuming and prone to errors
- some other criticism
  - corpora cannot provide negative evidence

Why use corpora in the first place?

- Chomsky's criticism
  - linguists study competence, not performance
- but
  - judgment data based on the intuition of the linguist himself are neither objective nor falsifiable
  - even performance lapses may reveal a lot of information about the underlying system producing both erroneous and 'correct' utterances
  - thus, corpus data are not irrelevant

Why use corpora in the first place?

- Chomsky's criticism
  - language is an unlimited system, corpora are finite
  - corpora are by definition incomplete: they will not contain possible sentences for linguistic reasons: because they are impolite, factually false, redundant
  - corpora are by definition biased because the likelihood of any sentence occurring in a corpus depends on its frequency, which in turn depends on non-linguistic features
    - *I live in Dayton, Ohio* vs. *I live in New York*

Why use corpora in the first place?

- Chomsky's criticism
  - language is an unlimited system, corpora are finite
- but
  - the sentences linguists dream up for their judgment data are often remote from anything native speakers actually say
  - the fact that the sentence I live in Dayton, Ohio is less frequent than I live in New York would not be relevant to corpus linguists since the observed frequencies of the two sentences correspond to the expected ones anyway

Why use corpora in the first place?

- Chomsky's criticism
  - 95% of all utterances are ungrammatical
  - thus, corpora are completely unrepresentative (since they overrepresent grammatical sentences)
  - thus, corpus data are irrelevant
- but
  - Chomsky has never provided any empirical evidence for this claim
  - the claim is simply false (cf. Labov 1969)

Why use corpora in the first place?

- Chomsky's criticism
  - corpora are unnecessary and you need judgment data anyway
  - the linguist (as a native speaker at least) can simply resort to his own intuition about the language in terms of generating
    - judgments
      - of grammaticality
      - of acceptability
      - of frequency
    - sentences

Why use corpora in the first place?

- Chomsky's criticism
  - corpora are unnecessary and you need judgment data anyway
- but
  - intuitions are often biased, fallible, and influenced by a variety of factors (just like performance!)
    - We received plans to kill Bill
    - *We received plans to kill each other
    - We received plans to kill me
      Chomsky's (1957:259) examples – which of these is (supposed to be) ungrammatical?
    - Labov (1975:89): "I have not been able to find any support among linguists or the general population for this judgment."

Why use corpora in the first place?

- Chomsky's criticism
  - corpora are unnecessary and you need judgment data anyway
- but
  - intuitions are often biased, fallible, and influenced by a variety of factors (just like performance!)
    - Grinder and Postal claimed that they could show interpretive semantics is wrong and generative semantics is right because of this judgment
    - *John didn't leave until midnight, but Bill did
    - Labov (1975:89): "Our own investigations […] failed to show any strong support for Grinder and Postal's position from linguists or other speakers"

Why use corpora in the first place?

- Chomsky's criticism
  - corpora are unnecessary and you need judgment data anyway
- but
  - this is not adequate scientific conduct: investigating a phenomenon and providing the data of the phenomenon at the same time is not objective, falsifiable, and valid
  - the judgments adduced by many linguists have not been elicited with all precautions required for complex experimental designs in psycholinguistics
  - only corpora provide data about speech and writing in authentic contexts

Why use corpora in the first place?

- Chomsky's criticism
  - corpus analyses are too time-consuming and prone to errors
    - data retrieval
    - data processing
    - data evaluation
- but
  - modern data processing techniques are not subject to these restrictions anymore
  - intuitive data are also prone to errors, but over and above that they are often also not explicit and objective
  - thus, corpus data have an advantage over intuitive data

Why use corpora in the first place?

- Others' criticism
  - corpora cannot provide negative evidence
  - if you do not find something in a corpus, this may just be because
    - the corpus is too small
    - people did not want to talk about X
  - no judgments about the impossibility of utterances are possible
- but
  - corpora can provide probabilistic negative evidence
  - if some linguistic element(s) occur(s) significantly less often than is statistically expected, this constitutes negative evidence

Why use corpora in the first place?

- Corpora can no only provide probabilistic negative evidence – corpora provide all sorts of probabilistic evidence
  - for example, with respect to language comprehension
    - accessing linguistic structures from the mental lexicon/grammar
    - disambiguation (of, say, ambiguous words and structures)
    - processing difficulties
  - for example, with respect to language production: accessing and assembling linguistic structures from the mental lexicon/grammar

Why use corpora in the first place?

- Corpora can provide information on what in legal circles is called the doctrine of ordinary meaning
- words not defined in a statute are to be understood in their ordinary meaning – but how does one determine that?
  - dictionaries are problematic
  - judges' intuitions are problematic
  - judges' decisions can be politically motivated
  - corpora can provide more objective and representative: what is/are the most frequent meaning(s) of an expression under consideration in a statute?
    - *use a gun*
    - *carry a gun*
    - *vehicles*