How to use corpus linguistics in sociolinguistics

Sociolinguistics

*The branch of linguistics

*It focuses on the effect which various social characteristics have on the language of individuals or speaker groups.

*It includes the major demographic categories of age, gender, social class, ethnicity, situational categories

*It focuses on the overall characteristics of language varieties: regional dialects, standard/non-standard varieties of a language, multilingualism, language policy, standardisation etc.

Corpus linguistics

*A methodological basis for doing linguistic research

*It comes to embody methodologies for linguistic description in which quantification is part of the research activity

- The first-generation written corpora - the Brown and Lancaster–Oslo/Bergen corpora.
- **Spoken corpora** - is a collection of speech data made accessible via a computer, containing at least transcriptions of speech but increasingly also audio and/or video files containing speech (the London–Lund Corpus (the spoken part of the British National Corpus).

3 Ways in which a sociolinguist may apply corpus-linguistic methods as an empirical basis

- 1 - the possibility of using one of the many existing corpora in a sociolinguistic study.
- 2 – to collect one's own data and create a new corpus for a specific sociolinguistic purpose
- 3 - the possibility of using 'corpus-inspired' methods to handle empirical data in a sociolinguistic study, without developing a fully-fledged corpus.

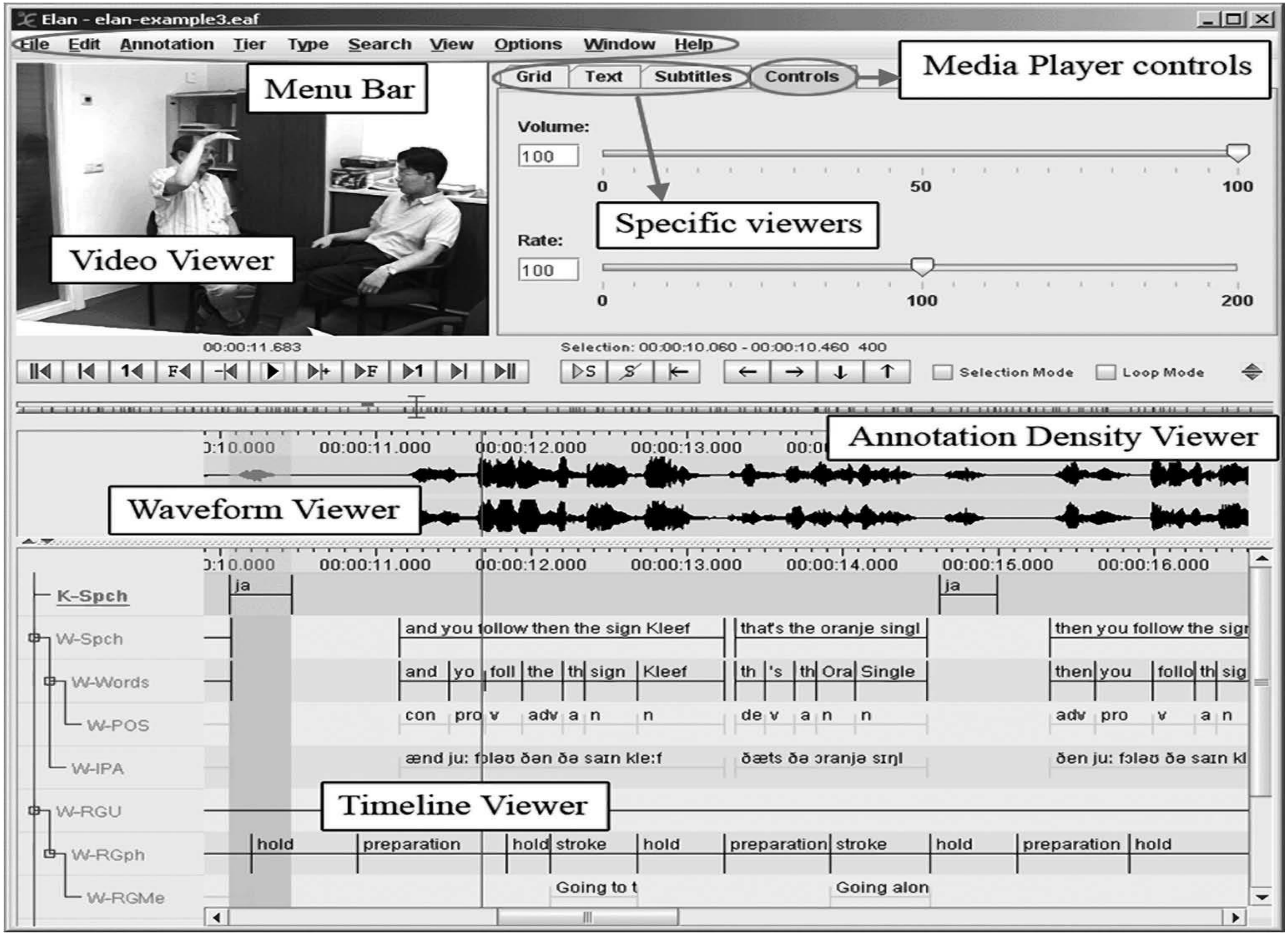
The corpus texts is annotated with linguistic information

- It may represent different levels of analysis:
- intonation,
- morphology,
- word class,
- syntactic structure,
- discourse structure
- speech act information

The corpus texts will be annotated with metadata

- the speakers,
- the situational context of the speech event,
- location,
- recording equipment,
- etc.

A flexible tool for transcription and annotation of spoken data is ELANT (the Eudico Linguistic Annotator),



* Types of spoken discourse in Spoken Dutch Corpus

- a. Spontaneous conversations (‘face-to-face’)
- b. Interviews with teachers of Dutch
- c. Spontaneous telephone dialogues (recorded via a switchboard)
- d. Spontaneous telephone dialogues (recorded on minidisk via a local interface)
- e. Simulated business negotiations
- f. Interviews/discussions/debates (broadcast)
- g. (political) Discussions/debates/meetings (non-broadcast)
- h. Lessons recorded in the classroom
- i. Live (e.g. sports) commentaries (broadcast)
- j. News reports/reportages (broadcast)
- k. News (broadcast)
- l. Commentaries/columns/reviews (broadcast)
- m. Ceremonious speeches/sermons
- n. Lectures/seminars
- o. Read speech

Advantages of spoken corpora like the BNC (British National Corpus) and the CGN (Spoken Dutch Corpus)

* It contains many speech genres, including the use of spoken language in formal settings like lectures, sermons and news broadcasts.

* It concerns their size and the possibility of efficiently processing large amounts of naturally occurring spoken data.

* E.g. The spoken part of the BNC includes 4,700 speakers representing many conversational settings.

the use of **like** as a discourse marker - a feature of white female speakers in late adolescence, and social class had no effect on the distribution of this marker

the use of the form **innit** - a feature of female adolescent speakers with an ethnic minority background and from the lowest socio-economic class.

LLC (London–Lund Corpus of spoken English)

* it exclusively involves speakers who are highly educated adults working in an academic environment

The spoken component of the BNC (British National Corpus)

contains ten million words of speech representing a variety of speech styles, including a substantial proportion of conversation in a variety of speaker groups.

A considerable weakness of both - the actual speech data, the sound files, are not available to the researcher.

A major concern in the development of speech corpora is how to treat non-standard and dialectal forms.

two ways of handling such variation

To produce a dialectal, orthophonic transcription which is as close to the pronunciation as possible

e.g. in the Nordic Dialect Corpus - **hvordan (how)**

To produce a transcription based on a written standard, but usually allowing for some variation.

e.g. in the Nordic Dialect Corpus –orthophonic realisations like **koless'n, kossj'n, korr**

PROBLEMS AND LIMITATIONS OF CORPUS LINGUISTIC METHOD

- 1 - The first objection concerns the accessibility of speech data, or lack thereof.
- 2 - some of the corpora that do have spoken data available are not suitable for detailed phonetic or phonological studies because the sound quality is too poor.
- 3 - there may be few instances of a particular variable in a corpus
- 4 - it may be problematic to rely fully on the judgements of the transcriber, since transcribers differ with respect to their interpretations of spoken data.
- 5 - some corpus data are only available as short snippets of sound and not as longer stretches of spoken dialogue.
- 6 - corpus linguistics relies heavily on the method of searching, and in many cases this is the only way to access the corpus data.
- 7 - corpus informants are generally anonymised. It prevents the researcher from accessing the in-depth background information.

Rules for corpus-based sociolinguistic research.

- 1 - Listen to the data
- 2 - Browse and search
- 3 – Precision and recall
- 4 - Distinguish form from function
- 5 - Manual work is needed
- 6 - Comparability of datasets

CORPUS-BASED SOCIOLINGUISTICS EXEMPLIFIED

* The project **Multicultural London English (MLE):** examines the role of ethnic-minority English in driving forward linguistic innovation at the levels of phonetics, grammar and discourse.

* **The Intonational Variation in English (IViE)** Corpus contains speech data and intonation transcriptions from nine urban dialects of British English. The data represent five different speaking styles.

* **Cheshire and Fox** (2009) study was/were variation in MLE.

* **McCarthy** (2002) focuses on the response tokens used in everyday conversation, basing the study on the British **CANCODE** (Cambridge and Nottingham Corpus of Discourse in English) corpus and the Cambridge–Cornell Corpus of Spoken North American English.