STATISTICAL METHODS IN COMPUTATIONAL LINGUISTICS – What can we expect?

Brigitte Krenn, ÖFAI, brigitte@ai.univie.ac.at

What Can Linguistics Expect From Statistics

- Augmentation of Competence Grammar with Performance Aspects of Language
- Data Orientation
- Processing Efficiency
- Robustness

Competence Performance

Language Competence is the knowledge of
an ideal adult native speaker-hearer to

- distinguish syntactically correct utterances from incorrect utterances,
- generate an infinite number of utterances from a restricted set of linguistic knowledge.

Grammaticality refers to competence.

Language Performance is

- reflected within actual language usage as can be found in corpora
- restricted by limitations of the human cognitive system such as restrictions of working memory, activation time, and access
- influenced by external factors such as restrictions due to a noisy channel, distractors, time limits

Acceptability refers to performance.

Competence Performance

Problems of Competence Grammars are

- Ambiguity
  - ↪ Large Search Space / Enumeration
  - ↪ Lack of Processing Efficiency

- Correct-False Dichotomy
  - ↪ Lack of Robustness
  - ↪ Only a Subset of Utterances is Considered

Lack of Performance Information

Computational Linguistics Needs for Competence-Performance Integration are due to

- Standard grammars have come to their limits.
- Applications need to be time efficient, robust and produce acceptable results.

Competence Ambiguity

Linguistic knowledge according to competence grammar is knowledge of ambiguous linguistic descriptions at different levels,

such as

- Word Level Syntactic Category
- Morphosyntactic Features
- Word Meaning
- Syntactic Structure
- Semantic Interpretation

Competence Ambiguity

Example:

Fruit flies like an arrow.

Word Level Ambiguities

Word Form	Syn. Cat.	Inflection	Word Meaning
fruit	noun verb	sg -3sg, intrans	product of plant or tree used as food cause to bear fruit
flies	noun verb	pl 3sg, intrans	insect move through the air
fruit flies	noun	pl	insect
like	verb preposition adjective	-3sg, trans – –	find pleasant in the manner of having the qualities of
an	article	sg	quantifier
arrow	noun	sg	$\rightarrow$

Competence Ambiguity

Syntactic Structure

(1)	Fruit	flies	like (an arrow)
	subject	head verb	modifier

(2)	Fruit flies	like	an arrow
	subject	head verb	argument

Competence Ambiguity

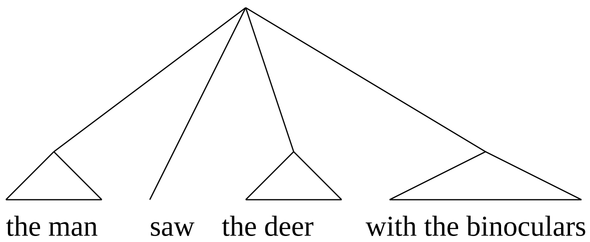
Two Classes of Phrase Level Ambiguities

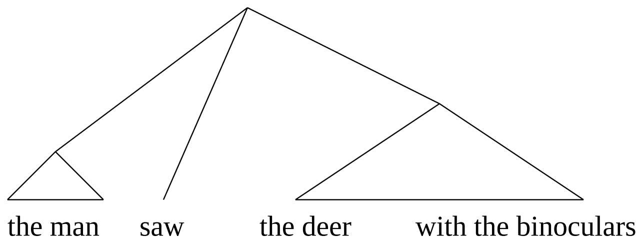
- Attachment Ambiguities
  - Prepositional Phrase Attachment
    The man saw the deer with the binoculars.
  - Relative Clause Attachment
    Someone shot the servant of the actress who was on the balcony.

- Scope Ambiguities
  - Focus Particles
    Bill only cleaned the bathroom.
  - Coordination
    Old men and women stayed at home.

Competence Ambiguity

Prepositional Phrase Attachment





Competence Ambiguity

Relative Clause Attachment

someone shot (the servant (of the actress who was on the balcony))

someone shot (the servant (of the actress) (who was on the balcony))

Syntactic Structure & Scope of *only*

Bill (only (cleaned the bathroom))

Bill ((only cleaned) (the bathroom))

Coordination

(old (men and women)) stayed at home

((old men) and women) stayed at home

Competence Correct False Dichotomy

Statistical language models are error tollerant.

Part-of-Speech Tagging

$$\operatorname{argmax}_{t_i} \prod P(t_i|t_{i-2}, t_{i-1}) P(w_i|t_i) \tag{1}$$

Combining Competence And Performance

Enhancement of Competence Grammars with Statistics

Stochastic Context-Free Grammar,
Stolcke, Segal 94, Mark et al.92,
Jelinek et al.92, Church 88

Stochastic Tree-Adjoining Grammar,
Schabes 92

Stochastic Unification-Based Grammar,
Briscoe, Carroll 93, Briscoe 94

Stochastic Dependency Grammar,
Eisner 96, Collins 96, Collins 97

Combining Competence And Performance

Integration of Competence and Performance Information within Statistical Models

Performance Data: Corpora

- Corpus data reflect language usage.

Competence Data: Corpus Annotations

- Annotations direct statistical generalizations.

Statistical Models: need to be designed to

- optimally combine relevant! linguistic information and occurrence frequencies.

Data Orientation

Corpus Types I: Text Type

Balanced Corpora comprise
different text genres
size proportional to
the distribution (relevance) of
a certain text type
e.g. the Brown Corpus,
the British National Corpus

Problem: the Notion of Relevance

Pyramidal Corpora range from
very large samples of a few representative genres to
small samples of a wide variety of genres

Problem: small amounts of data

Opportunistic Corpora: like
most corpora available, i.e.
"take what you can get!"

Data Orientation

Corpus Types I: Text Type

Desireable:

- sufficiently large samples
- various domains
- domains need to be indicated

Data Orientation

Corpus Types II: Kind of Annotation

Raw: text is

- tokenized and
- cleaned, e.g.
  control characters eliminated,
  text type, headlines, and paragraphs marked,
  cf. SGML mark-up

PoS-tagged: tokenized text is annotated with

- syntactic category at word level
  (part-of-speech PoS)

Data Orientation

Corpus Types II: Kind of Annotation

Treebanks: PoS-tagged text is annotated with

- sceletal syntactic structure
- a parse grammar is defined
- corpora are automatically parsed
- parse trees are selected and corrected by human annotators
- word strings for which no parse tree is automatically found are either omitted or manually annotated
- cf. the Penn Treebank I and II
  Marcus et al. 94

Linguistically Interpreted Corpora: aim at

- deliberate annotation of various kinds of linguistic information
- multi-stratal representation
- annotation is data-driven and descriptive
  Black et al. 96, Skut et al. 97

Data Orientation

Corpus Types III: Usage

Training Corpora: are

- used for statistical learning
- large
- typically annotated

Testing Corpora: are

- used for evaluation of statistical models
- small
- reference annotations

Testing on the Training Set to be avoided

Data Orientation

Some Examples:

Corpus	Susanne
Language	English
Size	128 000 words
Contents	Brown corpus
Annotation	parts-of-speech, syntactic structure

Corpus	Penn Treebank 1 + 2
Language	English
Contents	WSJ, ATIS, Brown
Annotation	parts-of-speech, syntactic structure 1 mio words (Penn 2) 1.6 mio words (Penn 1) PoS-tagged 1 mio words (Penn 1)

Corpus	BNC
Language	English
Size	100 mio words
Contents	text, speech
Annotation	parts-of-speech

Corpus	Frankfurter Rundschau
Language	German
Size	42 mio
Contents	newspaper
Annotation	parts-of-speech (currently 60 000 words), syntactic structure (currently 4 000 sentences)

What Can Linguistics Offer To Statistics

Corpus Annotation is combination of

- explicit,
- well understood linguistic information

with corpus data.

Means of Annotation

- Tagsets
- Annotation Schemes

What Can Linguistics Offer To Statistics

Classes of Tagsets

- Part-of-Speech Tags
- Phrasal Category Tags
- Grammatical Function Tags

Characteristics of Tagsets

- Moderate Size
- Hierarchical Organization
- Stratificational

Characteristics of Annotation Schemes

- Descriptive or Theory-Driven
- Cover all Data
- Multi-Stratal

What Can Linguistics Offer To Statistics

Typical Information in Part-of-Speech Tags

- Syntactic Category
- Selected Morphosyntactic Information
- Punctuation
- Foreign Material
- Extra Information
  e.g. multi-word, lexical trunc, abbreviation,
  pragmatic function, symbol, unknown

What Can Linguistics Offer To Statistics

Typical Treebank Architecture

- Set of Part-of-Speech Tags
- Set of Phrasal Category Tags
- Set of Function Tags
- Representation Formalism for Syntactic Structure

Two Major Representation Variants for Syntactic Structure

- Phrase Structure with Trace-Filler Mechanism
- Argument Structure
  (no distinction between local and non-local dependencies)

Treebanks And Linguistically Interpreted Corpora

Two Alternative Representations of Syntactic Structure

Phrase Structure	Argument (Dependency) Structure
Phenogrammatical Structure	Tectogrammatical Structure
Surface Order	Lexical Argument Structure

Two examples

The Penn Treebank & The NEGRA Annotation Scheme

Characteristics of the Penn Treebank	Characteristics of the NEGRA-Scheme
Phrase Structure Annotation	Argument Structure Annotation
Linear Order is represented within the tree structure by a trace-filler mechanism	Linear Order is represented by the word string
Labelled Bracketing	Unordered Trees

Treebanks And Linguistically Interpreted Corpora

Language Specific Features

English

- Rigid Word Order
- Restricted Inventory of Discontinuous Constituency

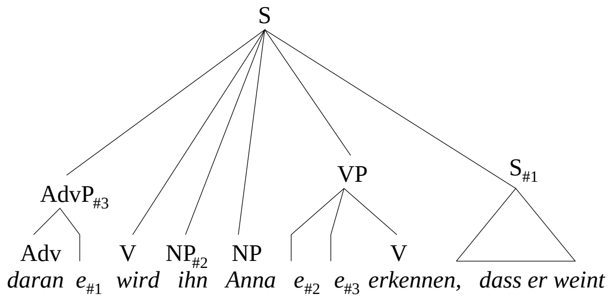
German

- Continuum of Local and Non-Local Dependencies
- Massive Occurrence of Discontinuous Constituency
  e.g. *daran wird ihn Anna erkennen, dass er weint*
- Word Order Variation Influenced by Many Factors
  e.g. category, syntactic function, focus

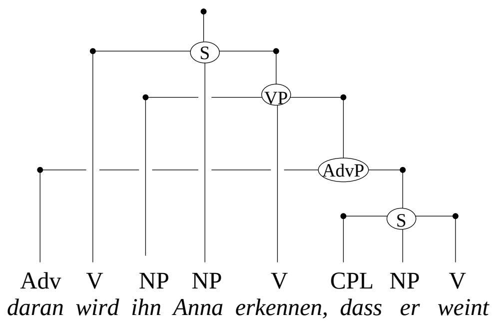
- Continuum of Grammaticality of Word Permutation
  ihn wird Anna daran erkennen, dass er weint
  Anna wird ihn daran erkennen, dass er weint
  erkennen wird ihn Anna daran, dass er weint
  daran, dass er weint, wird ihn Anna erkennen
  dass er weint, daran wird ihn Anna erkennen

Treebanks and Linguistically Interpreted Corpora

Phrase Structure Representation



Argument Structure Representation



Treebanks And Linguistically Interpreted Corpora

Representation Formalism: Penn Treebank 2

- Phrase Structure: Ordered Trees + Trace-Filler Mechanism
  Labelled Bracketing
- Grammatical Functions: Bracket Labels
- Syntactic Categories: Bracket Labels

Representation Formalism: NEGRA-Scheme

- Argument Structure: Unordered Trees
- Grammatical Functions: Edge Labels
- Syntactic Categories: Node Labels

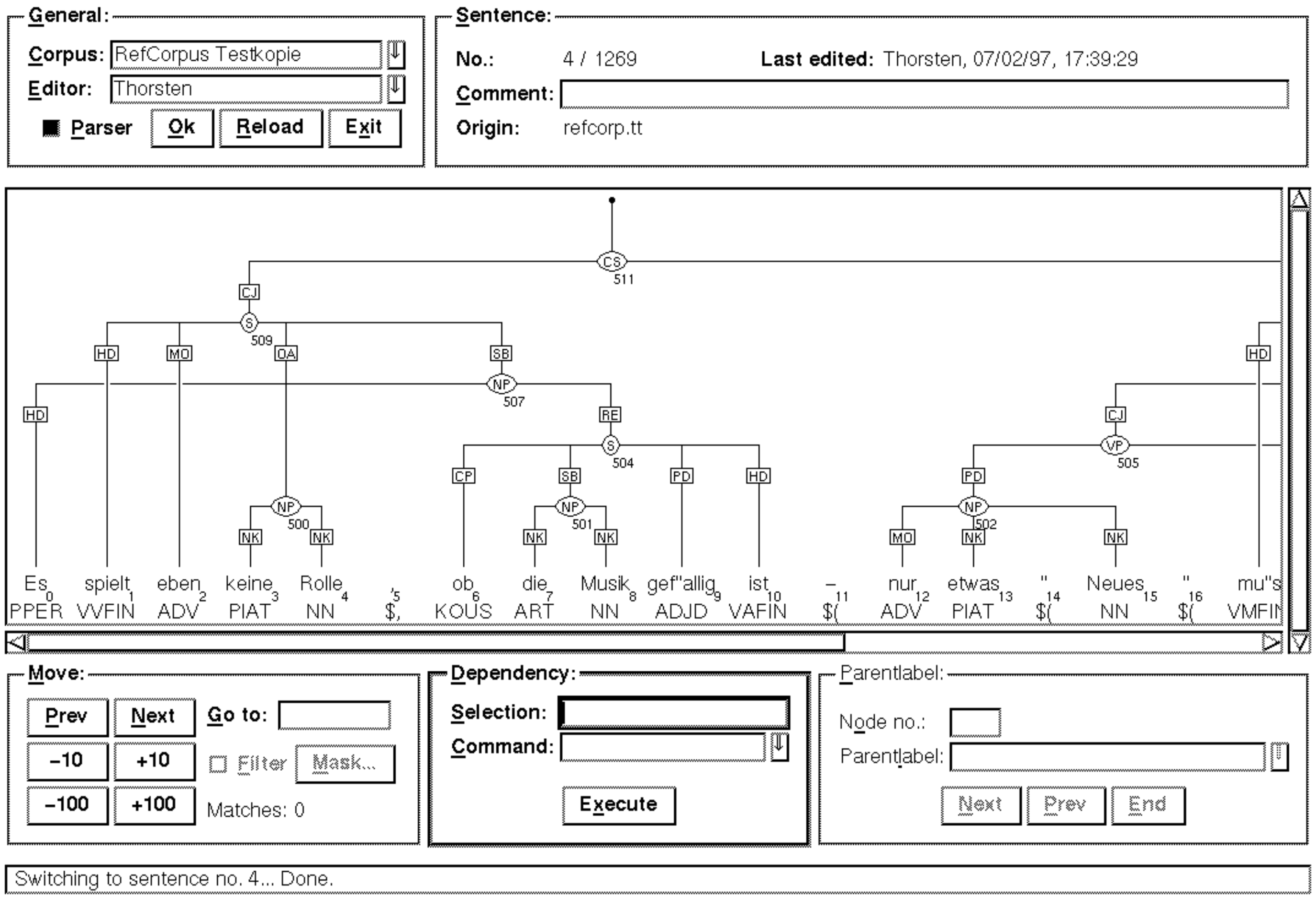
Treebanks And Linguistically Interpreted Corpora

Argument Structure Annotation

Requirements on the Annotation Tool: User Interface

- Graphical Representation
  to display unordered trees
- Mouse Input
  can be efficient

Screen Dump of the Annotation Tool



Treebanks And Linguistically Interpreted Corpora

NEGRA Annotation Automation: Boot-Strapping

- Fully Manual
- Manual: Phrase Boundaries and Syntactic Category
  Automatic: Assignment of Grammatical Functions
- Manual: Phrase Components
  Automatic: Syntactic Category, Grammatical Functions
- Automatic: Basic Phrases
- Automatic: Partial or Complete Parses

Tagging Grammatical Functions

The Technique Used: Standard Part-of-Speech Tagging

$$\text{argmax}_g \prod P(t_i|t_{i-2}, t_{i-1}) P(g_i|t_i) \tag{2}$$