Language

Carola Eschenbach, Christopher Habel, Wolfgang Menzel

Department Informatik
Universität Hamburg

5. November 2008

Language

- Language
- Syntax (Wolfgang Menzel)
- Meaning (Carola Eschenbach)

Language ...

- ... is an extremely complex phenomenon,
  but seemingly easy to learn
- ... is ruled by regularities, but full of exceptions
- ... is fairly stable, but extremely flexible
  - universal means to communicate arbitrary content
  - adaptation to the needs and capabilities of the hearer/speaker
  - establishes social cohesion

Language ...

... is studied from different perspectives

- language acquisition
- language learning
- grammar
- social behaviour
- language change
- human language processing
- computational processing
- ...

→ different research goals, different research methodologies

Language ...

... is studied on different levels of analysis

- phonetics: How to pronounce words?
- phonology: How to distinguish meaning?
- morphology: How to produce (new) words and word forms?
- syntax: How to combine word forms into sentences?
- lexical semantics: How to establish the meaning of words?
- compositional semantics: How to establish the meaning of a sentence?
- discourse: How sentences are combined into text or dialogue?
- pragmatics: What's the communicative function of an utterance in a specific context: speaker and hearer, situation, ...?
- ...

→ abstraction and simplification is necessary

Is language special?

- yes, because there is no other species
  - which developed a similar complex and universal system of communication and
  - which can learn our language to the same degree of proficiency
- no, because
  - it seems to work on the same kind of neural substrate as other cognitive faculties do

Language Processing

- natural language processing by machines falls way short of the human model with respect to
  - coverage
  - learnability, adaptability
  - robustness
- Cognitive modelling
  - might help to overcome some of these deficiencies
  - is essential for deep processing of natural language
  - even with a strong focus on applications
    - What do humans do when they understand / interpret / produce an utterance?
    - What kind of processing capabilities of their communication partners do humans expect?

Language

- Language
- Syntax (Wolfgang Menzel)
- Meaning (Carola Eschenbach)

Syntax

- Why syntax?
- Representations
- Grammars
- Parsing
- Preferences

Why syntax?

- Syntax as a linguist's darling.
- Semantics as the poor sister of syntax.
- Why is this so?
- looking for meaning but finding structures
  - easy access to empirical data
  - more regularities: high potential for general descriptions
  - models better scale up

Different perspectives

The linguist's perspective: understanding language

- What's the difference between a natural and an artificial language?
- Why natural languages can be learned by humans, but artificial ones can't?
  - by any child / in a surprisingly short time
- How do humans process natural language?
  - almost effortless / in real time / in a highly robust manner

Different perspectives

The linguist's perspective (cont.)

- What's the interplay between different aspects of language in order to facilitate communication?
  - lexicon / morphology / syntax / semantics / pragmatics
  - specifically: how syntax mediates between form and meaning

Different perspectives

The computer scientist's perspective:
language understanding (and production) systems

- How to make a machine ...
  ... to understand natural language content
  ... to express content by means of natural language
- Which other useful tasks can be accomplished without actually "understanding" language?
  - hyphenation
  - spell checking
  - text-to-speech synthesis
  - grammar checking
  - ...

Different perspectives

The computer scientist's perspective (cont.)

- If content access is the goal: is syntax really necessary?
  - Schank: Conceptual dependency
  - bag-of-words approaches
- but: ignoring syntax is no solution either

Why syntax?

- common misconception:
  syntax is (only) about the correctness of utterances
- but:
  checking for correctness is only one particularly important empirical technique
- syntax is ...
  ... about the (underlying) structures
  ... the interface to semantics

Why syntax?

- syntax facilitates language understanding
  - key to any ambitious natural language application
- e.g. argument assignment: who does what to whom?
  - English: ordering, configurational language

    The man$_{\text{subject}}$ *bought his wife*$_{\text{indirect object}}$ *a book*$_{\text{direct object}}$.
  - German: (semi-)free word order language (scrambling)
    $\rightarrow$ arguments are distinguished by means of case

    Der Mann$_{\text{nom}}$ *kaufte seiner Frau*$_{\text{gen,dat}}$ *ein Buch*$_{\text{nom,acc}}$.
    Seiner Frau$_{\text{gen,dat}}$ *kaufte der Mann*$_{\text{nom}}$ *ein Buch*$_{\text{nom,acc}}$.
    Ein Buch$_{\text{nom,acc}}$ *kaufte der Mann*$_{\text{nom}}$ *seiner Frau*$_{\text{gen,dat}}$.
    Der Mann$_{\text{nom}}$ *kaufte ein Buch*$_{\text{nom,acc}}$ *seiner Frau*$_{\text{gen,dat}}$.
    ...

Why syntax?

- e.g. negation: what's possibly negated?

  The man did not buy his wife a book.
  The man did not buy his wife a book.
  The man did not buy his wife a book.
  The man did not buy his wife a book.

- e.g. pronouns: what's a potential candidate for reference?

  John$_i$ cannot begin, before he$_i$ arrives.
  Before he$_i$ arrives, John$_i$ cannot begin.
  Before John$_i$ arrives, he$_i$ cannot begin.
  **He$_i$ cannot begin, before John$_i$ arrives.*

Why syntax?

- immediate applications
  - dialogue systems
  - content extraction from text
  - machine translation
  - report generation
- other syntax driven applications
  - language modelling: predicting the next word
    - problem: long-range dependencies
      → structured language models
  - quality assessment
    - grammar checking
    - language tutoring

Why syntax?

- example: disambiguation for text-to-speech synthesis

  The girls will read the paper.

  The girls have read the paper.

  Will the girls read the paper?

  Have any men of good will read the paper?

  Have the executors of the will read the paper?

  Have the girls who will be away next week read the paper?

  Please have the girls read the paper.

  Have the girls read the paper?

Jackendoff & LSA

A central question

Is (human) language processing rule driven (generic) or instance / example based?

- Do we parse or do we remember?
- Do we generate or do we reproduce?

The problem

- instance-based processing cannot explain language learning and innovative use of language
- but:
  there are many non-productive idiocrasies: idioms, metaphoric use, etc.:
  to catch a cold
  to reach for a star
- language production is habitual: stereotypical utterances
  you are welcome
  ladies and gentlemen
- outright exceptions
  I enjoyed this meeting, said he.
  That's a real problem, believe you me.

Did you know?

I sometimes read a book.
Sometimes I read a book.

I rarely read a book.
Rarely do I read a book.

I always read a book.
**Always ... .*

The problem

Q: What's the most important English rule?

A: Almost every rule is about 90% valid.

from Kenneth Beare at about.com
English as a 2nd language

The problem

- language processing is a mixture of rule driven and instance-based procedures
- c.f. transfer-based translation vs. example-based translation
- The engineer's concern:
  - How to combine these approaches?
  - What's the proper balance between them?

Language: Syntax

- Why syntax?
- Syntactic Representations
- Grammars
- Parsing
- Preferences

Representations

- The poor man's syntax: shallow structures
  - part-of-speech tags: noun, auxiliary, full verb, adjective

    The girls will$_{AUX}$ *read the paper.*
    Have$_{AUX}$ *any men of good will*$_{N}$ *read the paper?*

  - syntactic segments: chunks

    Will [$_{NP}$ *the girls] read the paper?*
    Have [$_{NP}$ *the executors] [*$_{PP}$ *of the will] read the paper?*

Representations

- full syntactic descriptions are (at least) hierarchical structures (trees)

  [$_{S}$ *Have [*$_{NP}$ *the girls*
  [$_{S}$ *who [*$_{VP}$ *will be away [*$_{NP}$ *next week]]]]*
  [$_{VP}$ *read [*$_{NP}$*the paper?]]]*

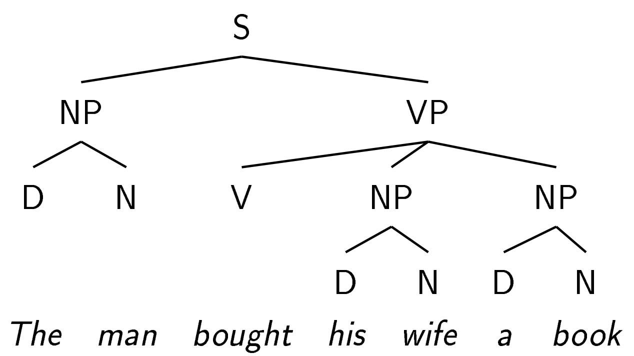
  [$_{S}$ *Please [*$_{VP}$*have [*$_{NP}$ *the girls]] [*$_{VP}$*read [*$_{NP}$ *the paper]]]]].*

  [$_{S}$ *Have [*$_{NP}$ *the girls] [*$_{VP}$ *read [*$_{NP}$ *the paper?]]]*

- two different approaches for syntactic representations:
  - phrase structure grammars
  - dependency grammars
- emphasize different aspects of syntax
- can partly be transformed into each other

Phrase structure grammars

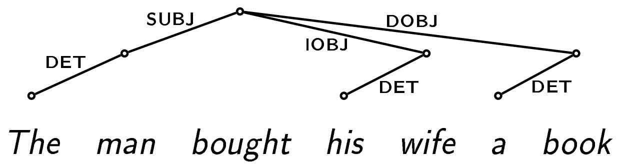
- typed, recursive grouping of word forms



- empirical basis
  - distributional analysis
  - substitution of partial trees
- particularly suited for configurational languages (fixed word order)

Dependency grammars

- functional perspective: modification of words by others with respect to a particular syntactic function (Subject, Object, ...)



- easy mapping to the semantic level: thematic roles
- particularly suited for non-configurational (free word order) languages
  - slavonic languages, German, Dutch, Japanese, ...
- regularly structured and finite space of partial structures
  - well suited for some machine learning approaches and constraint satisfaction procedures
- allows to model non-projective (not properly nested) structures

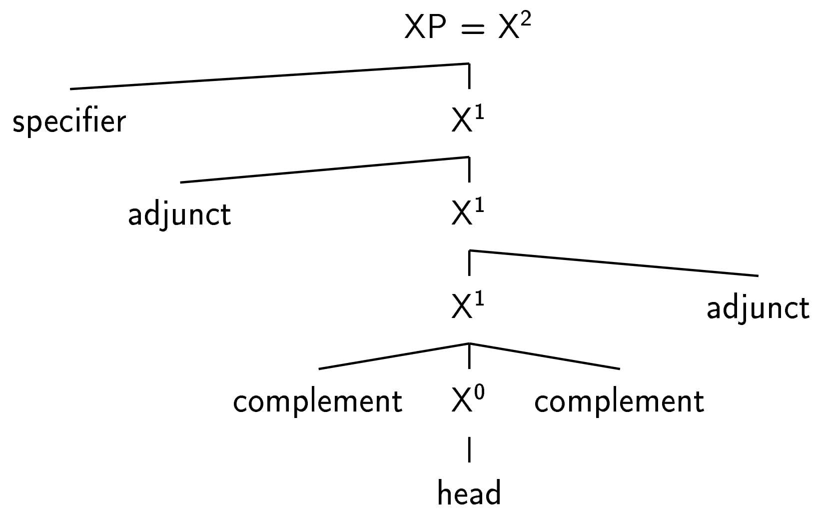
Treebanks

- huge collections of real-life sentences
- manually/semi-automatically annotated with tree structures
- many different annotation conventions
  - kind of structures: phrase vs. dependency structure
  - bracketing: projective vs. non-projective
  - theoretical assumptions: X-bar-theory, DP-modelling, tecto-grammatical structures, ...)
  - depth: deeply nested vs. shallow structures (e.g. compound words, phrases)
  - granularity of POS tags
  - granularity of labels
  - use of empty nodes
- used for training and testing of NLP systems

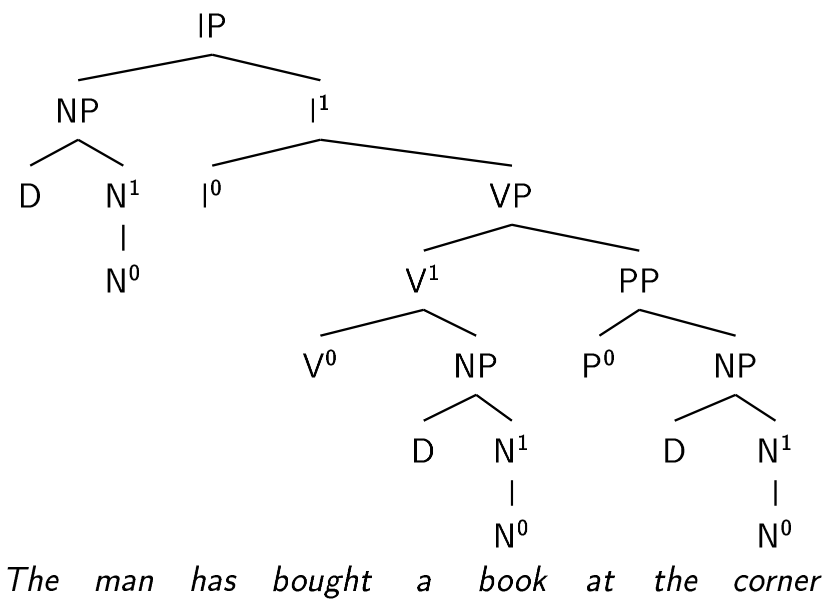
X-bar-theory

- Jackendoff (1979)
- universal tree pattern which ...
  - ... restricts the space of possible structural descriptions
  - ... presumably underlies all languages of the world
  - ... reflects the human language faculty
  - ... is meant to explain, why every human child can learn an arbitrary human language
- strong type restrictions: phrasal/lexical nodes
- allows a local description of various syntactic phenomena: government, projection, agreement, ...
- requires movement operations
  - → not well suited for machine learning approaches
  - → not used in treebank annotations

X-bar-theory

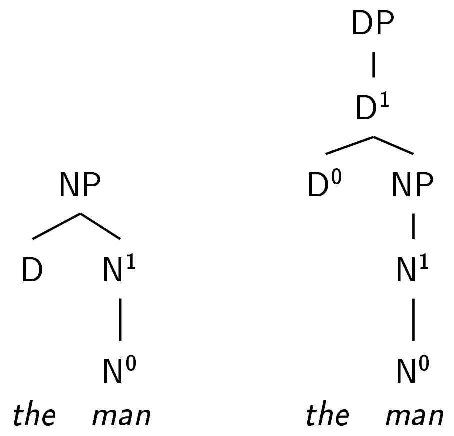


X-bar-theory



DP-modelling

- radical application of the X-bar restrictions to the NP



Language: Syntax

- Why syntax?
- Representations
- Grammars
- Parsing
- Preferences

Grammars

- generative approaches:
  - rules are used as building blocks to construct trees
  - rules can be extracted from the tree annotation

S	→	NP VP
VP	→	V NP
VP	→	VP PP
NP	→	D N
...		
N	→	book
N	→	man
D	→	the
...		

- → context-free grammar (CFG, Chomsky 1953)
- rules can be interpreted as rewrite possibilities

Grammars

- X-bar constraints would reduce the number of possible rule-types considerably:

XP	$\longrightarrow$	SpecX X^1
X^1	$\longrightarrow$	X YP
X^1	$\longrightarrow$	YP X
X^1	$\longrightarrow$	X^0
X^1	$\longrightarrow$	X^0 YP
X^1	$\longrightarrow$	X^0 YP ZP

- ... but require the additional movement operator

Grammars

- finer grained modelling: unification-based approaches
- additional features + unification requirements

S	$\longrightarrow$	NP[Per,Num] VP[Per,Num]
VP[Per,Num]	$\longrightarrow$	V[Per,Num,Case] NP[Case,X,Y]
NP[Case,Num,Gen]	$\longrightarrow$	D[Case,Num,Gen] N[Case,Num,Gen]
...		

- complex categories: recursively embedded feature structures
  $\longrightarrow$ potentially infinitely many
  $\longrightarrow$ not well suited for machine learning approaches
- unification can be used to construct arbitrary structural descriptions: parse trees, semantic forms, ...

Grammars

- major drawback of rule-(and unification)-based grammars: usually fairly low coverage even for large grammars
- typical examples for German
  - partial Parser (Wauschkuhn 1996)
    - 56.5% on newspaper text
  - Gepard: based on a unification grammar (Langer 2001):
    - 33.51% on newspaper text
    - up to 66% on testsuites (better lexical coverage, shorter and less ambiguous sentences)

Grammars

- an alternative: constraint-based approaches
- constraint is a very general notion
- one needs to distinguish: constraints over ...
  - ... complex feature structures
    $\longrightarrow$ constraint-based formalisms:
    - formalism for logical deduction
    - e.g. head-driven phrase-structure grammar (HPSG)
  - ... elements of a dependency structure
    $\longrightarrow$ constraint grammars:
    - formalism for constraint satisfaction
    - e.g. (weighted) constraint dependency grammar (CDG, WCDG, ...)

Language: Syntax

- Why syntax?
- Representations
- Grammars
- Parsing
- Preferences

Parsing

- given a sentence (sequence of word forms and interpunction symbols)
- determine one or possibly several parse trees
- problem: ambiguity

  He bought the book with his wife.
- local ambiguity: alternative rules can be applied to the same data

  VP → V NP
  He bought the book with his wife.

  VP → VP PP
  He bought the book with his wife.
- global ambiguity: several structures for a sentence

Parsing strategies

- rules can be applied in two different ways:
  - replace the left-hand side of a rule with its right-hand side: top down
  - replace the right-hand side of a rule with its left-hand side: bottom up
- alternatives can be considered in a different order
  - all in parallel: breadth-first
  - one at a time: depth-first
- the sentence can be processed
  - left-to-right
  - right-to-left

Parsing

highly simplified example: top-down, depth-first, left-to-right

S							
NP	VP						
D	N	VP					backtracking
PersP	VP						
he	VP						
	V	NP					
	bought	NP					
		D	N				
		the	N				
			book				backtracking
		NP	PP				
		D	N	PP			
		the	N	PP			
			book	PP			
				P	NP		
				with	NP		
					D	N	
					his	N	
						wife	backtracking
	VP	PP					
...							
he	bought	the	book	with	his	wife	

Parsing

- non-deterministic algorithm: choice points available
- naive approaches require exponential effort
- reuse of partial structures: cubic effort

  → chart parsing
- serious problem: broad coverage grammars lead to highly ambiguous output

Parsing

- *Hinter dem Betrug werden die gleichen Täter vermutet, die während der vergangenen Tage in Griechenland gefälschte Banknoten in Umlauf brachten.*
- two unification-based parser for German:
  - Paragram (Kuhn and Rohrer 1997):
    - LFG-grammar
    - 92 readings
  - Gepard (Langer 2001)):
    - special unification-based grammar
    - 220 readings
    - average ambiguity on newspaper text: 78 readings (average sentence length 11.43 words)
- extrem case: $6.4875 \cdot 10^{22}$ readings for another German sentence (Block 1995)

Parsing

- What went wrong?
- Writing a grammar is struggeling against Zipfean Law
  - roughly: the probability of an item is inversely proportional to its rank
  - holds for almost all language phenomena: phones, lexical items, rules
  - there are few items which are frequent, but very many which are rare
- modeling the first 90% is easy, but catching the rest becomes increasingly difficult

Parsing

- for parsing a sentence, several rules need to be applied, probably including rare ones, which have not been modelled yet

  → low coverage
- increasing the coverage, means writing more rules
- more rules will increase the degree of local ambiguity
  → more global ambiguity

Parsing

- the alternative perspective:
  constraint grammars for dependency models
  - instead of using generative rewrite rules ...
  - ... constraints on the wellformedness of structural descriptions are specified
- all structures are admitted unless explicitly ruled out
- default reasoning: the last remaining structure survives
  - full coverage: the parser never fails!
  - but: usually no full disambiguation can be achieved

Language: Syntax

- Why syntax?
- Representations
- Grammars
- Parsing
- Preferences

Preferences

- context-free grammars are based on a first order axiomatisation
  - enumeration of solutions is possible ...
  - ... but no comparison of solutions
- no ranking of hypotheses according to plausibility available
- no selection among the potential readings of a sentence can be performed
- solution: definition of a weighting scheme over rules/constraints
- two different approaches
  - Optimality theory
  - Weighted constraint dependency grammar

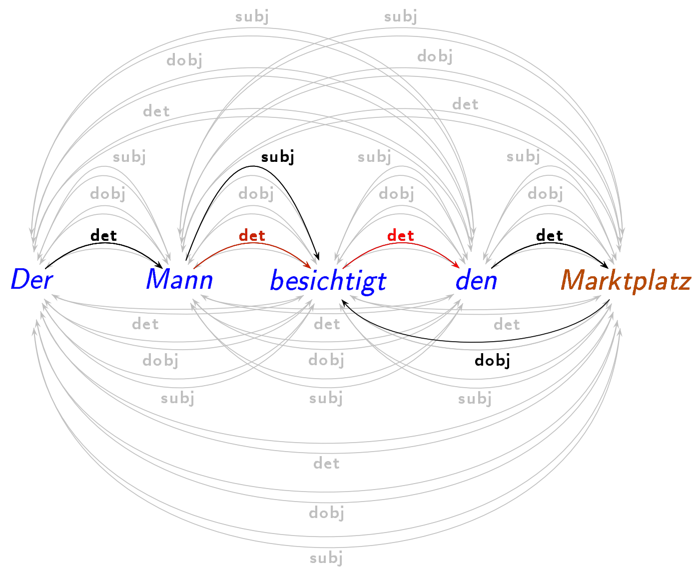
Optimality theory

- Prince and Smolensky (1993)
- grammar with a context-free backbone
- constraints are ...
  - ... local within a rule
  - ... ordered according to their relative strength in a hierarchy
- claim:
  - context-free backbone and constraints are universal
  - ranking is language specific and needs to be learned
- the grammar assigns the structure which only violates the least important constraints
- parsing becomes an optimization problem
  $\longrightarrow$ full disambiguation
- but: no broad coverage models available so far

Weighted constraint dependency grammar

- Menzel 1995, Schröder (2001), Foth (2005)
- no generative component at all
- constraints on admissible dependency structures
- constraints are weighted
- weights are combined multiplicatively
- heuristic decision procedures try to determine the optimal structural description

Transformation-based parsing



Weighted constraint dependency grammar

- constraints are used to include predictions of external components
  - tagger, chunker, supertagger, PP-attacher, attachment predictor
- results (structural/labelled accuracy):
  - WCDG with tagger only: 89.7% / 87.9%
  - WCDG with full set of predictors 92.5% / 91.1%

Preferences

- can preferences be learned from data?
- different classes of machine learning approaches:
  1. predicting the structure vs. predicting the parser actions
  2. generative vs. discriminative learning

Generative models

- learning: estimation of probability distributions
- deciding on the maximum posterior probability

$$\arg\max_t p(t|s)$$

- posterior probability cannot be estimated directly
- reformulation as

$$p(t|s) = \frac{p(t)p(s|t)}{p(s)}$$

- task: finding the optimal predictor for the input
- well suited for phrase-structure grammars

Generative models

- simplest case: treebank grammars
  - extract the rules from a treebank
  - estimate their probability $p(lhs \longrightarrow rhs|lhs)$
- fairly poor results (Charniak 1996)
  - sentences up to max. 40 word forms:
  - labelled recall = 80.4%, labelled precision = 78.8%
- What went wrong?

1. treebank grammars generalize poorly
   - the treebank is far too small
   - Penn-treebank: relatively flat structures
   - 40000 sentences $\longrightarrow$ 10605 rules
   - 3943 occurring only once in the corpus!

Generative models

2. rule-probabilities do not capture the relevant information
   - rule application probability also depends on lexical relationships

   $\longrightarrow$ lexicalized probabilities (Charniak 2000, Collins 1999))

   p$(lhs \longrightarrow rhs|lhs, head(rhs))$

   p$(lhs \longrightarrow rhs|lhs, head(rhs), head(mother(rhs)))$

   ...

   - results (labelled recall / precision, Charniak 2000):
     - sentences ($l \leq 40$ words): 91.0% / 91.0%
     - all sentences: 89.6% / 89.5%
   - to compare with a treebank grammar:
     - sentences ($l \leq 40$ words): 80.4% / 78.8%

Discriminative models

- directly approximating $p(t|s)$ without a generative model
- finding the optimal class boundary or function approximation
- learning: modifying a high-dimensional function to optimally approximate the target
- examples: neural networks, support vector machines

Discriminative models

- application to dependency parsing:
  MST-parser (McDonald 2006)
- on-line learning of a weighting function for local dependency hypotheses
- maximum spanning tree-search ($O(n^2)$) based on local scores
- transformation-based search based for higher order dependencies
- best parser on the CoNLL 2006 shared task:
  - 91.5% structural accuracy for English
  - 90.4% / 87.3% structural / labelled accuracy for German
- Can we do even better?

Discriminative models

- replication of the experiments with another annotation standard for German
- almost the same results:
  - 90.5% / 87.5% without interpunction
  - 91.9% / 89.3% with interpunction
- combination with WCDG as another predictor
  (results with interpunction)

	without MST	with MST
WCDG + tagger	89.7% / 87.9%	93.0% / 91.8%
WCDG + all predictors	92.5% / 91.1%	93.9% / 92.6%

Summary

- syntax is an important factor of human language comprehension
- syntactic structures are important for many NLP applications
- local ambiguity in broad coverage grammars make 1st-order axiomatizations intractable
- preferential reasoning is required to rank hypotheses according to plausibility
- parsing becomes an optimization problem
- rule-based and trained empirical knowledge can be combined successfully

Outlook

- Are we hitting the ceiling?
- What's next?
- human language communication is situated
  - environment
  - background knowledge
  - intentions

  He ate the sandwich with his wife.
- more semantics?
- more world knowledge?
- better user models?
- better machine learning techniques?
- higher-level inference techniques?