Cross-modal Interaction between Language and Vision

Wolfgang Menzel

Universität Hamburg, Fachbereich Informatik

Cross-modal Interaction between Language and Vision

- Cross-modal comprehension in the human model
- Weighted Constraint Dependency Grammar
- Predictive parsing
- Parsing in the visual world paradigm
- What's next?

Cross-modal comprehension in the human model

- establishes a bidirectional interaction
  - language guides visual attention
  - vision helps to process speech and language

Cross-modal comprehension in the human model

- establishes a bidirectional interaction
  - language guides visual attention
  - vision helps to process speech and language
- exhibits a high degree of robustness: able to deal with
  - cross-modal conflict
  - incomplete evidence
  - uncertainty
  - structural and referential ambiguity

Cross-modal comprehension in the human model

- establishes a bidirectional interaction
  - language guides visual attention
  - vision helps to process speech and language
- exhibits a high degree of robustness: able to deal with
  - cross-modal conflict
  - incomplete evidence
  - uncertainty
  - structural and referential ambiguity

Cross-modal comprehension in the human model

- takes timely decisions in dynamic environments
  - utterances unfold over time $\rightarrow$ incremental analysis
  - vision is attention driven $\rightarrow$ stepwise refinement
  - revision of intermediate interpretations might be necessary

Cross-modal comprehension in the human model

- takes timely decisions in dynamic environments
  - utterances unfold over time $\rightarrow$ incremental analysis
  - vision is attention driven $\rightarrow$ stepwise refinement
  - revision of intermediate interpretations might be necessary
- carries out an extremely rapid information fusion
  - largely based on the anticipation of upcoming stimuli

Cross-modal comprehension in the human model

- takes timely decisions in dynamic environments
  - utterances unfold over time $\rightarrow$ incremental analysis
  - vision is attention driven $\rightarrow$ stepwise refinement
  - revision of intermediate interpretations might be necessary
- carries out an extremely rapid information fusion
  - largely based on the anticipation of upcoming stimuli
- is usually studied in the visual world paradigm
  - looking at a picture while listening to a related utterance
  - monitoring the eye movements to study the time course of reference resolution

Weighted Constraint Dependency Grammar

- conditions for dependency edges or combinations of them
- individual penalties for the severity of constraint violations
- searching for the optimal combination of edges
- most successful solution method: frobbing
  - start with an initial structure (educated guess)
  - find the most severe constraint violation
  - choose an alternative attachment, edge label, or morpho-syntactic reading to repair it
- heuristic taboo search: non-optimal results have to be expected

Weighted Constraint Dependency Grammar

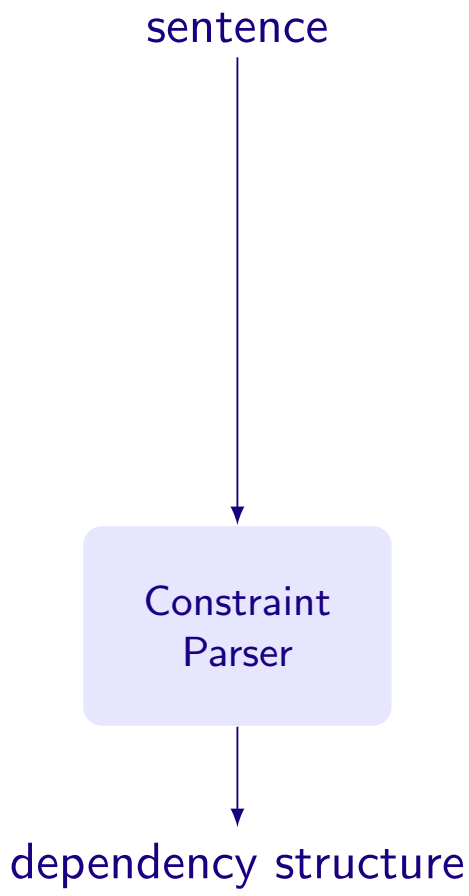
- high accuracy due to
  - preferential reasoning
  - the use of a huge dictionary ($>$ 1 mio inflectional forms)
  - the integration of shallow predictive components
- but accuracy is not the whole story

Weighted Constraint Dependency Grammar

- predictors are trained independently on corpus data
- predictors are unreliable
- predictions might contradict each other

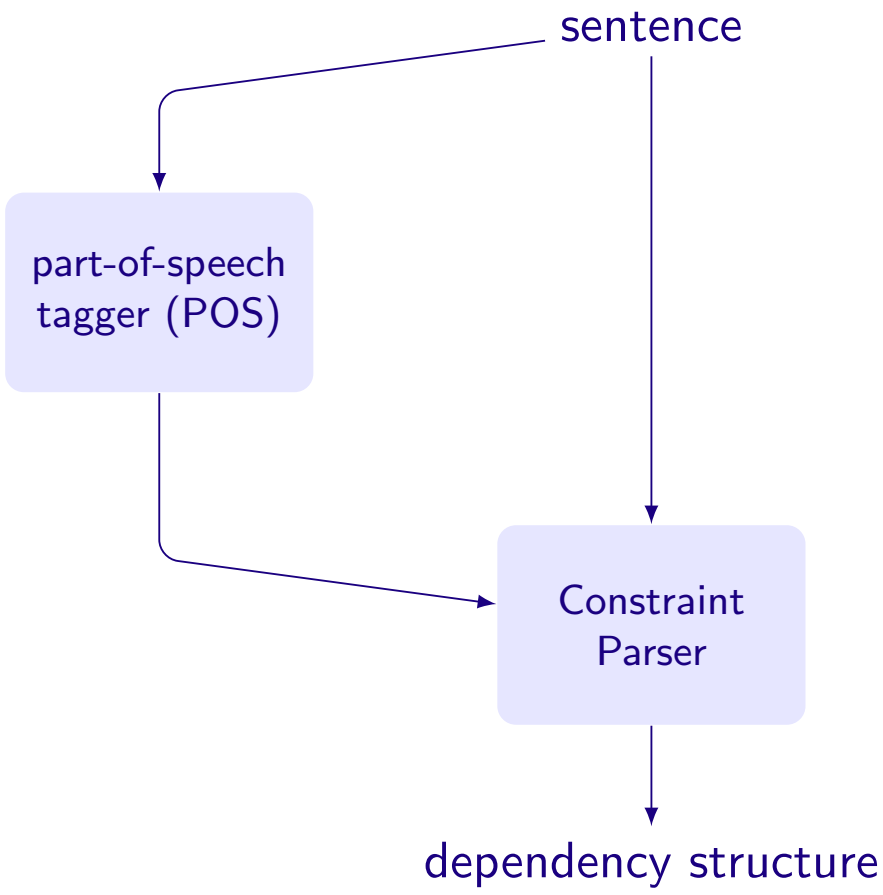
Weighted Constraint Dependency Grammar

- predictors are trained independently on corpus data
- predictors are unreliable
- predictions might contradict each other



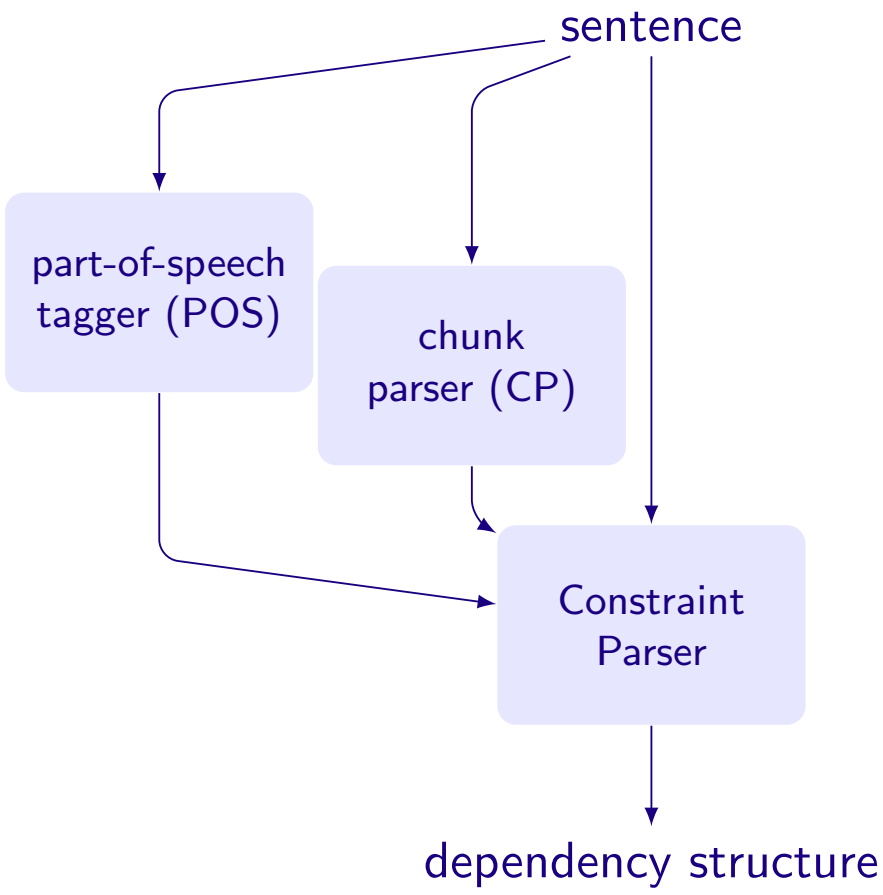
Weighted Constraint Dependency Grammar

- predictors are trained independently on corpus data
- predictors are unreliable
- predictions might contradict each other



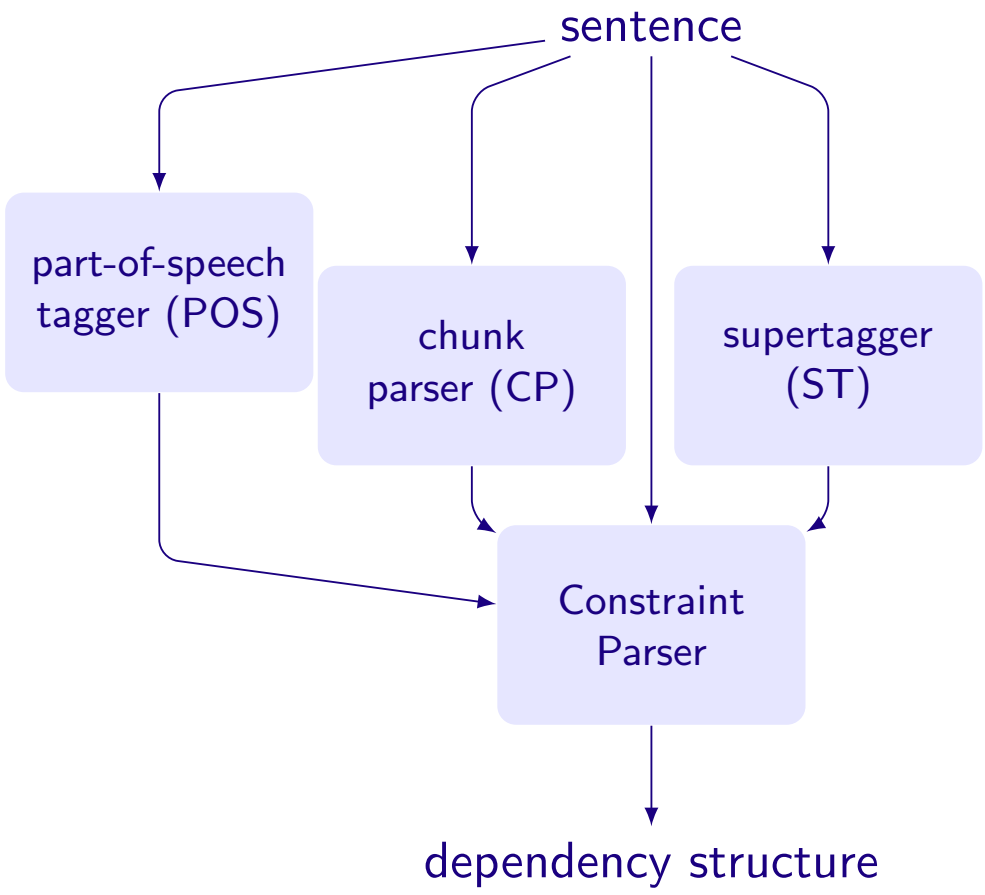
Weighted Constraint Dependency Grammar

- predictors are trained independently on corpus data
- predictors are unreliable
- predictions might contradict each other



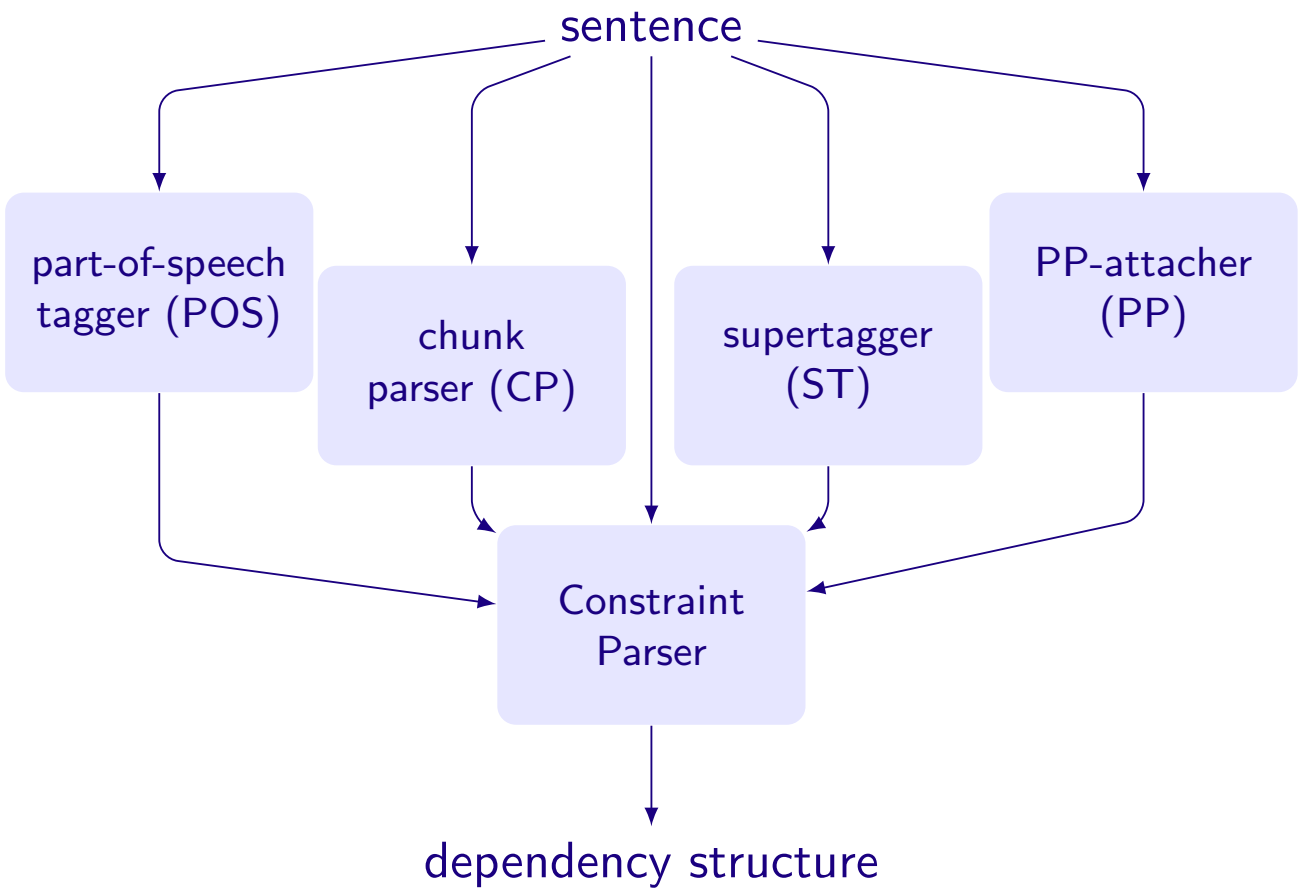
Weighted Constraint Dependency Grammar

- predictors are trained independently on corpus data
- predictors are unreliable
- predictions might contradict each other



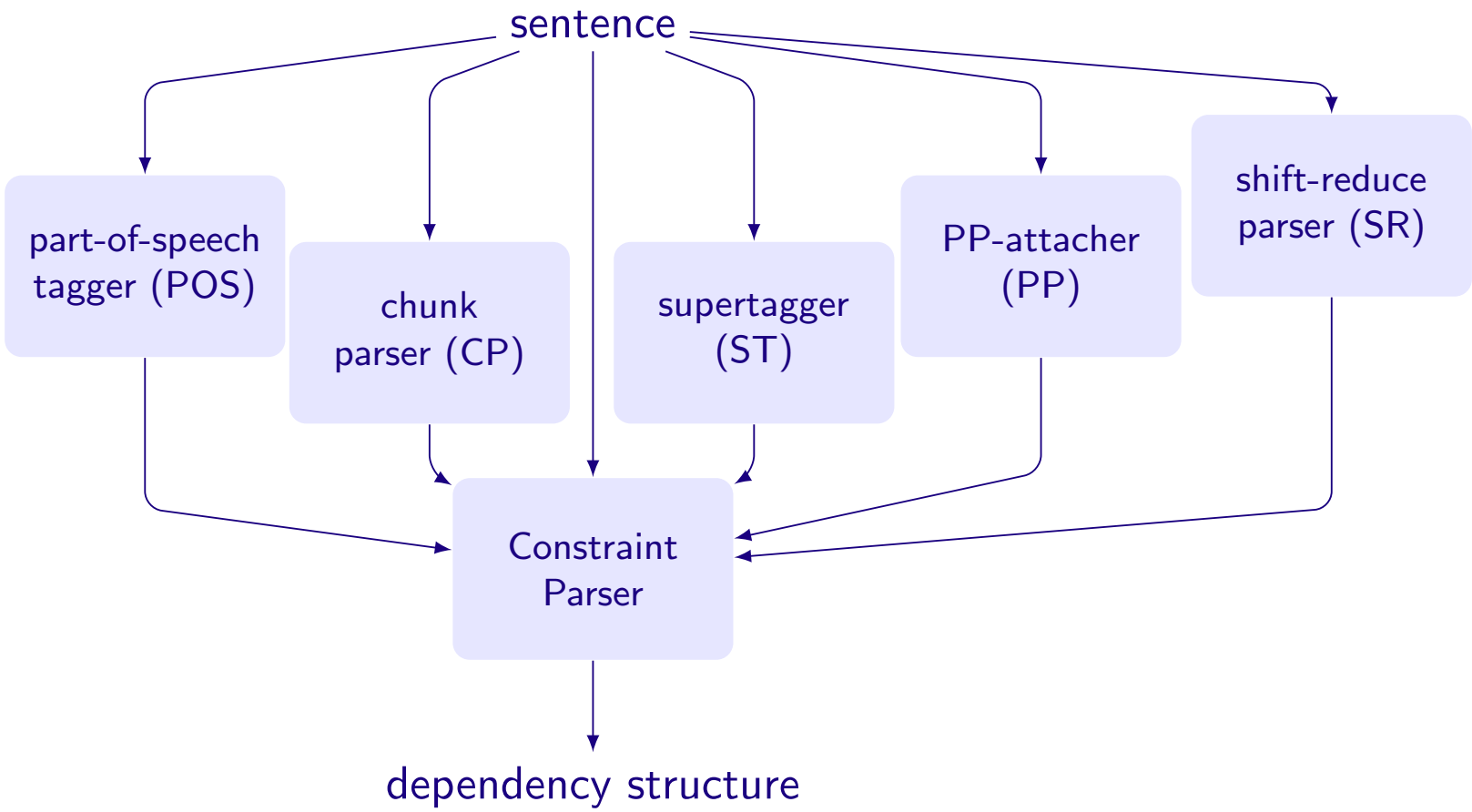
Weighted Constraint Dependency Grammar

- predictors are trained independently on corpus data
- predictors are unreliable
- predictions might contradict each other



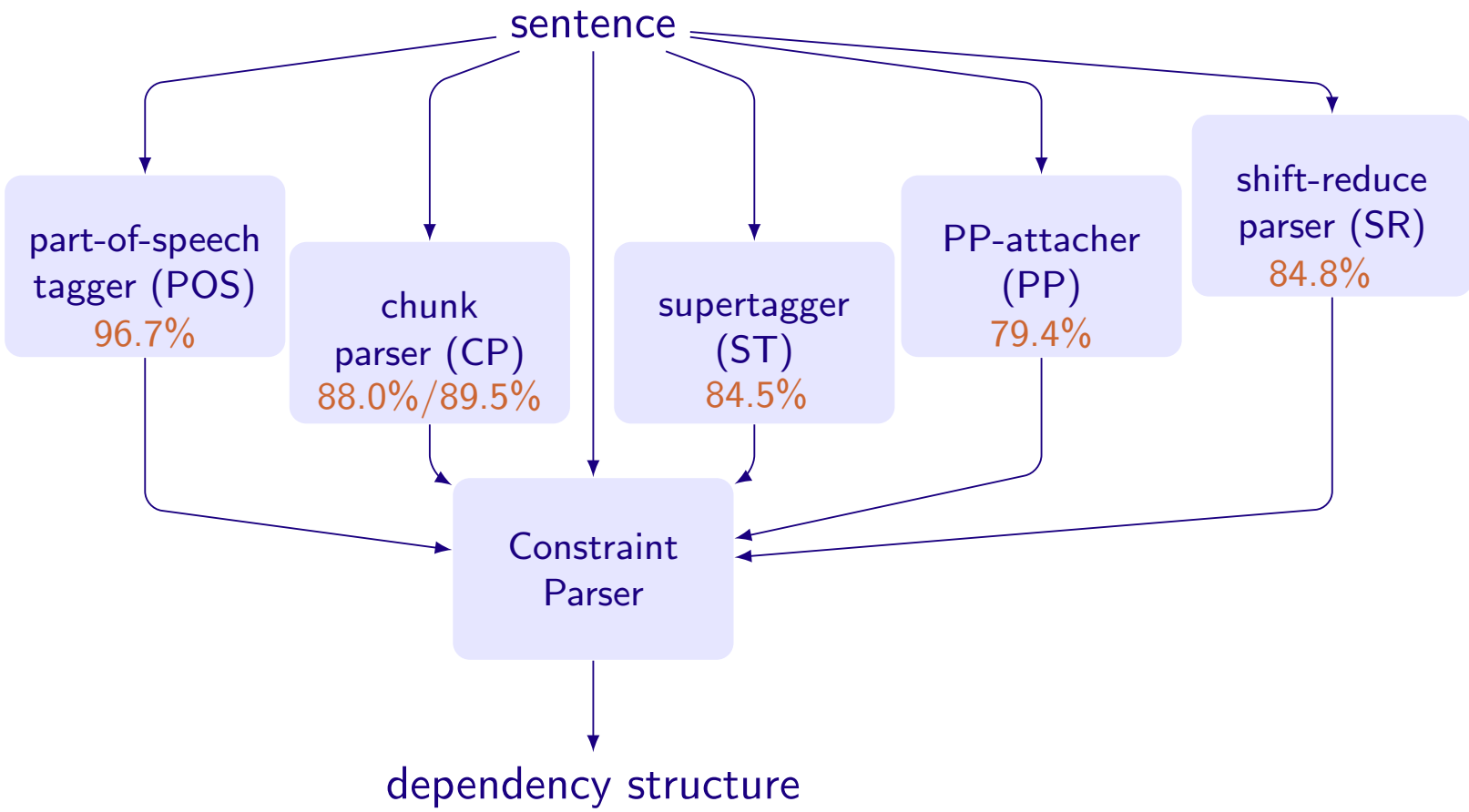
Weighted Constraint Dependency Grammar

- predictors are trained independently on corpus data
- predictors are unreliable
- predictions might contradict each other



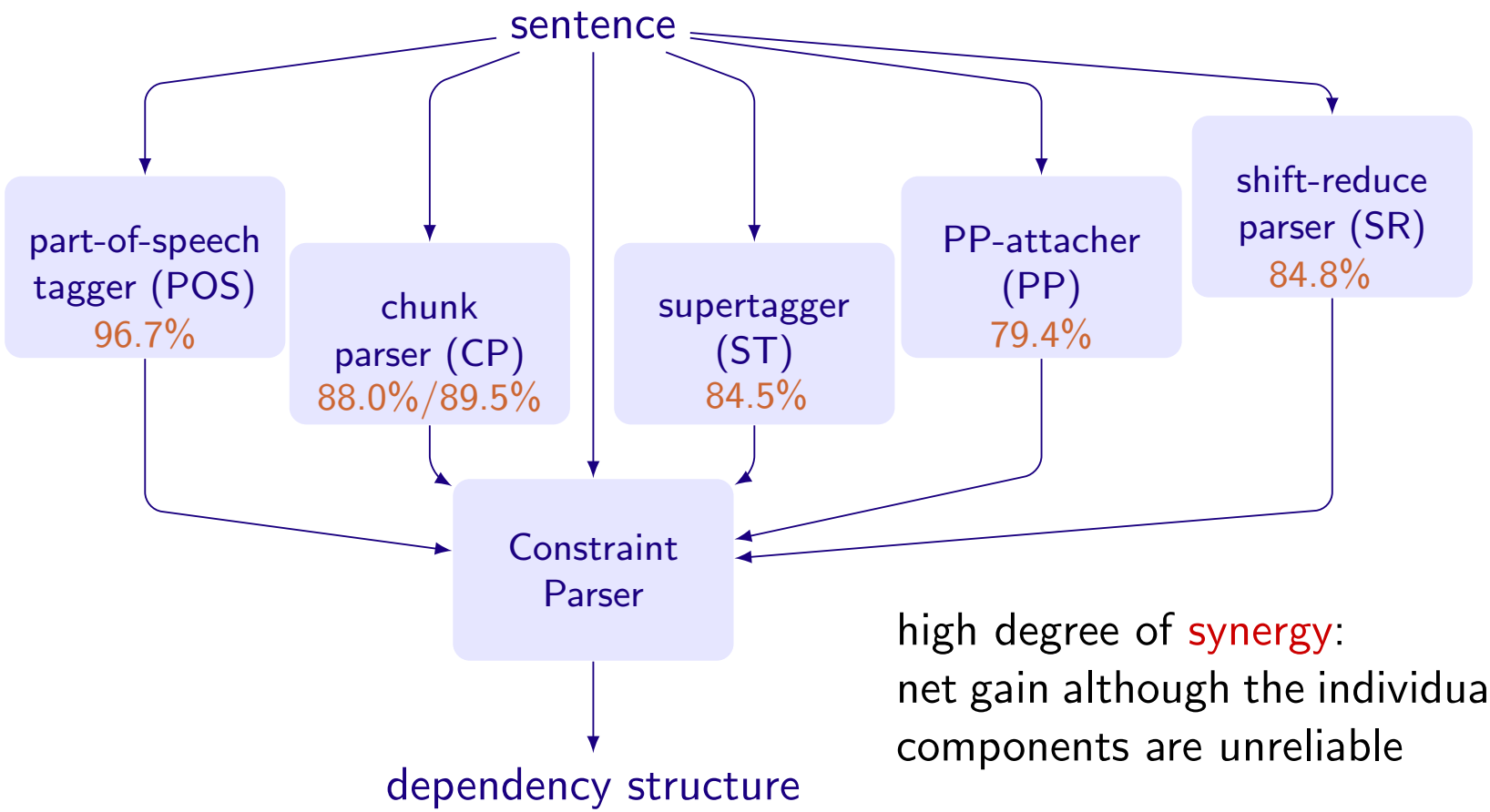
Weighted Constraint Dependency Grammar

- predictors are trained independently on corpus data
- predictors are unreliable
- predictions might contradict each other



Weighted Constraint Dependency Grammar

- predictors are trained independently on corpus data
- predictors are unreliable
- predictions might contradict each other



high degree of synergy:
net gain although the individual components are unreliable

Weighted Constraint Dependency Grammar

- constraints provide alterative input channels
  - (part of) the structure can be (weakly) predefined

Weighted Constraint Dependency Grammar

- constraints provide alterative input channels
  - (part of) the structure can be (weakly) predefined
- strong anytime behaviour: interruptable and decisive
  - being able to return the most promising interpretation upon request
  - at the price of revisions and reanalysis effort

Weighted Constraint Dependency Grammar

- constraints provide alterative input channels
  - (part of) the structure can be (weakly) predefined
- strong anytime behaviour: interruptable and decisive
  - being able to return the most promising interpretation upon request
  - at the price of revisions and reanalysis effort
- robustness: ability to deal with internal conflicts
  - grammar rules are not consistent (e.g. preferences)
  - the utterance and its context may contradict
  - incomplete utterances usually are usually ungrammatical
  - actual continuations may violate expectations

Weighted Constraint Dependency Grammar

- constraints provide alterative input channels
  - (part of) the structure can be (weakly) predefined
- strong anytime behaviour: interruptable and decisive
  - being able to return the most promising interpretation upon request
  - at the price of revisions and reanalysis effort
- robustness: ability to deal with internal conflicts
  - grammar rules are not consistent (e.g. preferences)
  - the utterance and its context may contradict
  - incomplete utterances usually are usually ungrammatical
  - actual continuations may violate expectations
- high potential for error diagnosis
  - constraint violations indicate deviations from the norm

Weighted Constraint Dependency Grammar

- ability to disambiguate on multiple (related) description layers in parallel

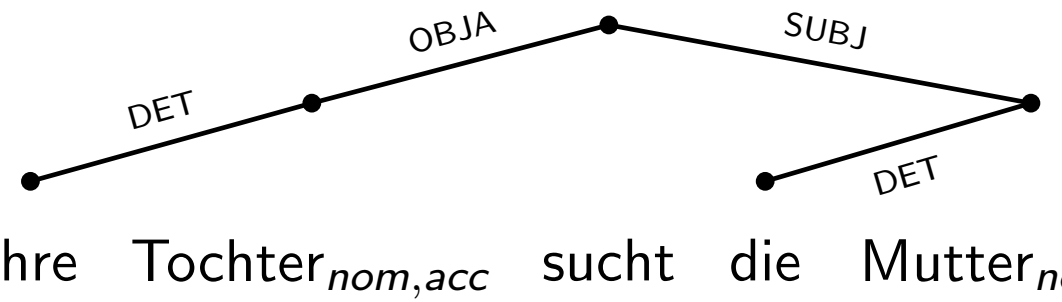
Weighted Constraint Dependency Grammar

- ability to disambiguate on multiple (related) description layers in parallel

Ihre Tochter$_{nom,acc}$ sucht die Mutter$_{nom,acc}$

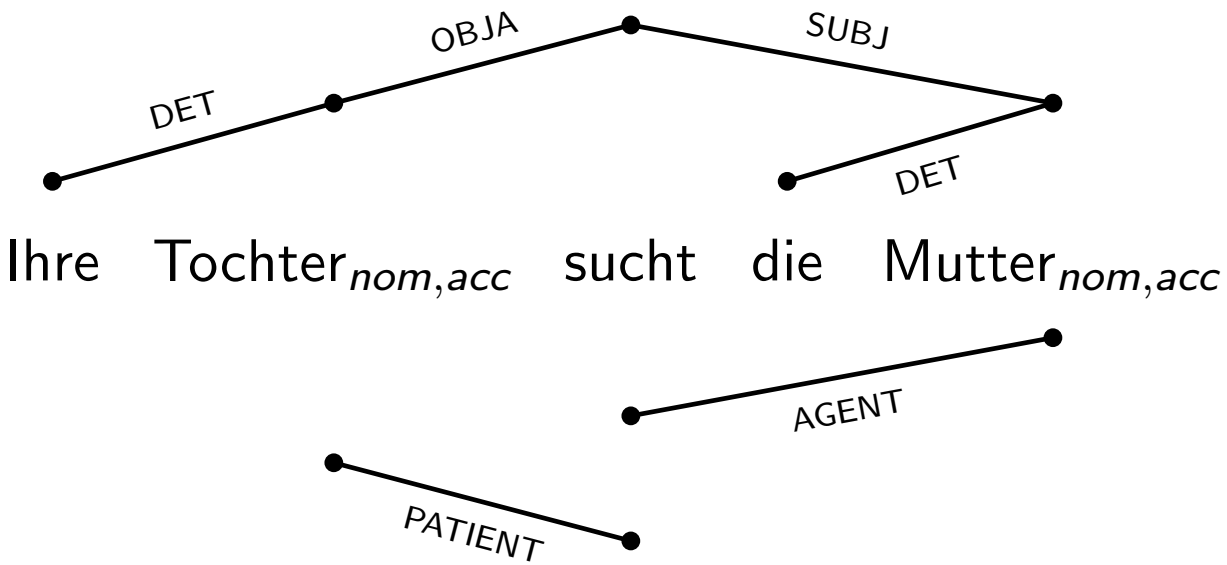
Weighted Constraint Dependency Grammar

- ability to disambiguate on multiple (related) description layers in parallel



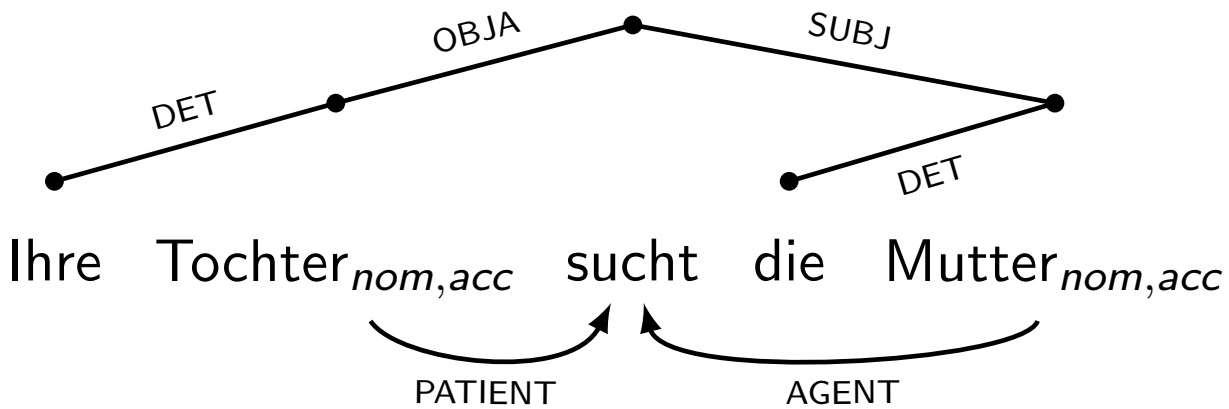
Weighted Constraint Dependency Grammar

- ability to disambiguate on multiple (related) description layers in parallel



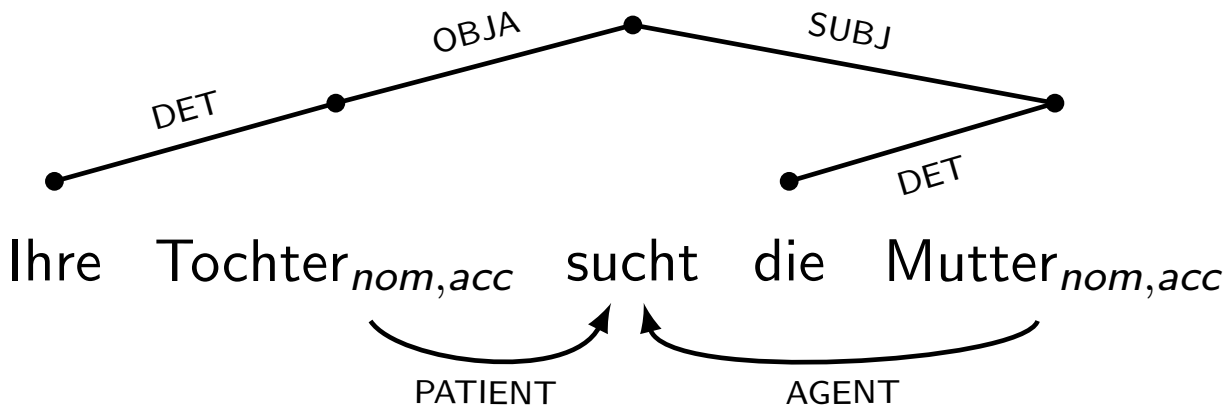
Weighted Constraint Dependency Grammar

- ability to disambiguate on multiple (related) description layers in parallel



Weighted Constraint Dependency Grammar

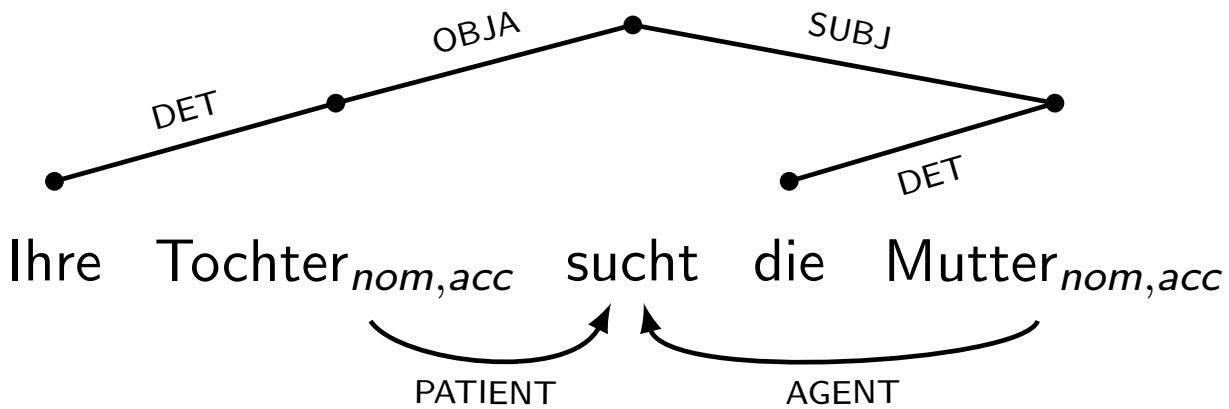
- ability to disambiguate on multiple (related) description layers in parallel



- weak coupling: mapping of representational layers with weighted constraints

Weighted Constraint Dependency Grammar

- ability to disambiguate on multiple (related) description layers in parallel



- weak coupling: mapping of representational layers with weighted constraints
- sensitivity to extra-linguistic contextual influences
  - constraints are truly relational: interface between description layers can be made bidirectional

Contributions and contributors

Ingo Schröder (2000)

first protype system,
experiments with various solution strategies

Kilian Foth (2006)

hybrid parsing, broad coverage grammar,
state-of-the-art accuracy for German

Patrick McCrae (2010)

cross-modal parsing with simulated visual input,
effects of underspecified and uncertain evidence

Contributions and contributors

Niels Beuck (forthcoming)
incremental predictive parsing, using
virtual nodes to integrate upcoming lexical items

Arne Köhn (2013)
multi-threaded implementation,
incremental parsing in push mode

Christopher Baumgärtner (2013)
guiding visual attention with linguistic predictions,
parsing in dynamically changing contexts

Predictive parsing

- naive incremental processing attaches a word as soon as possible
  - $\rightarrow$ eager processing, but not eager enough
  - attachment is pending until the attachment point becomes available
  - results in a fragmented (less informative) interpretation

Predictive parsing

- naive incremental processing attaches a word as soon as possible
  - $\rightarrow$ eager processing, but not eager enough
  - attachment is pending until the attachment point becomes available
  - results in a fragmented (less informative) interpretation
- predictive parsing collects pending attachments using virtual nodes
  - virtual nodes can be integrated into the current structure
  - allow to build a connected (more informative) structure

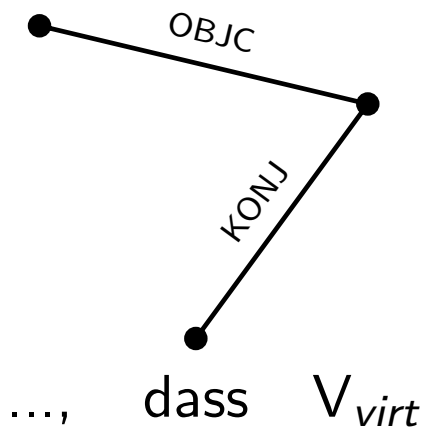
Predictive parsing

- naive incremental processing attaches a word as soon as possible
  - $\rightarrow$ eager processing, but not eager enough
  - attachment is pending until the attachment point becomes available
  - results in a fragmented (less informative) interpretation
- predictive parsing collects pending attachments using virtual nodes
  - virtual nodes can be integrated into the current structure
  - allow to build a connected (more informative) structure

..., dass

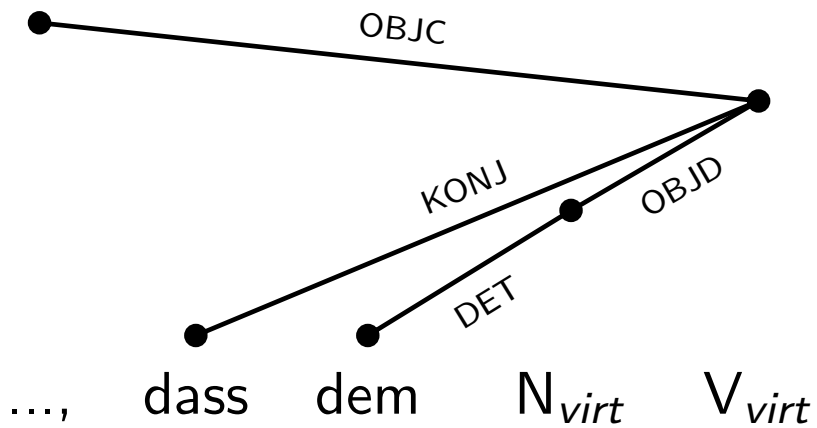
Predictive parsing

- naive incremental processing attaches a word as soon as possible
  - $\rightarrow$ eager processing, but not eager enough
  - attachment is pending until the attachment point becomes available
  - results in a fragmented (less informative) interpretation
- predictive parsing collects pending attachments using virtual nodes
  - virtual nodes can be integrated into the current structure
  - allow to build a connected (more informative) structure



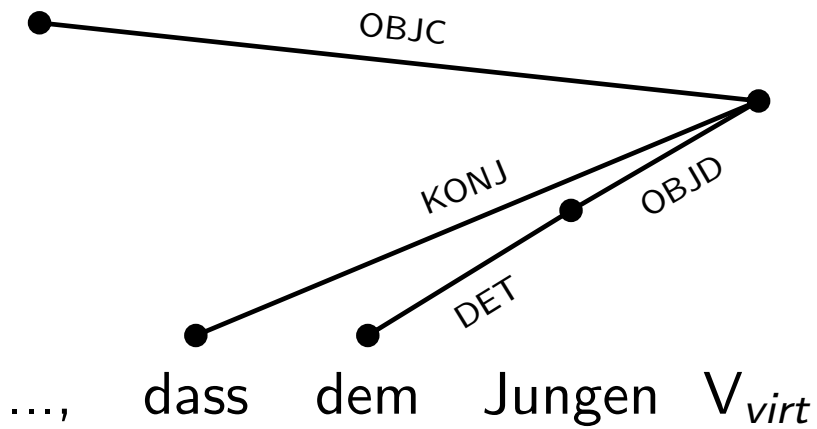
Predictive parsing

- naive incremental processing attaches a word as soon as possible
  - $\rightarrow$ eager processing, but not eager enough
  - attachment is pending until the attachment point becomes available
  - results in a fragmented (less informative) interpretation
- predictive parsing collects pending attachments using virtual nodes
  - virtual nodes can be integrated into the current structure
  - allow to build a connected (more informative) structure



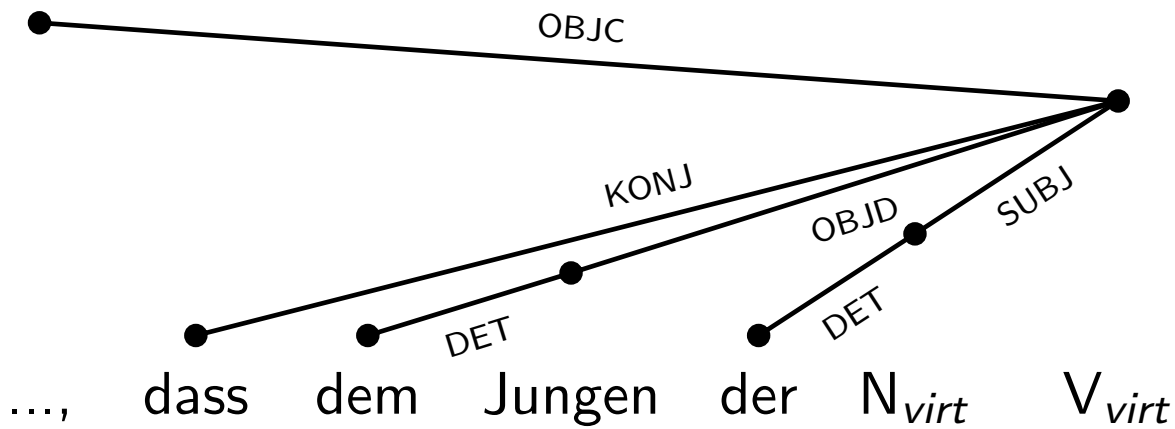
Predictive parsing

- naive incremental processing attaches a word as soon as possible
  - $\rightarrow$ eager processing, but not eager enough
  - attachment is pending until the attachment point becomes available
  - results in a fragmented (less informative) interpretation
- predictive parsing collects pending attachments using virtual nodes
  - virtual nodes can be integrated into the current structure
  - allow to build a connected (more informative) structure



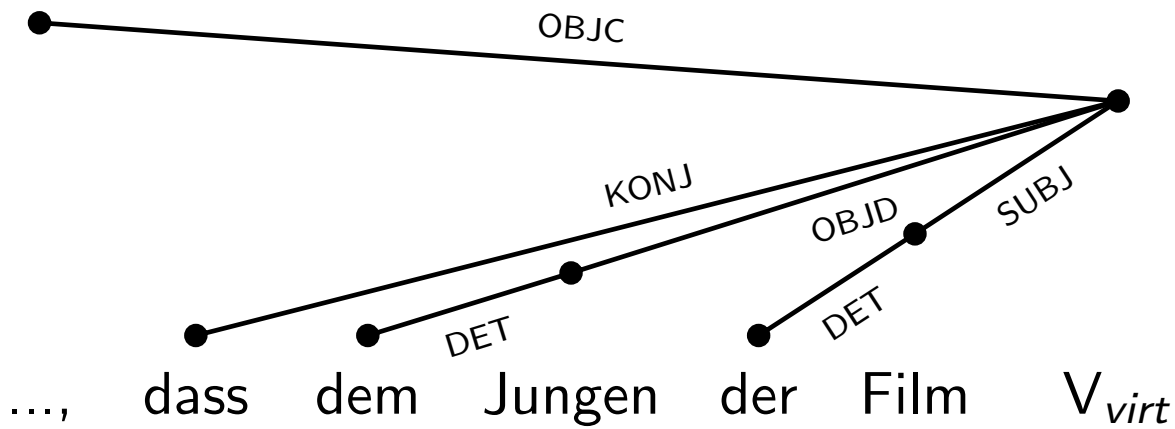
Predictive parsing

- naive incremental processing attaches a word as soon as possible
  - $\rightarrow$ eager processing, but not eager enough
  - attachment is pending until the attachment point becomes available
  - results in a fragmented (less informative) interpretation
- predictive parsing collects pending attachments using virtual nodes
  - virtual nodes can be integrated into the current structure
  - allow to build a connected (more informative) structure



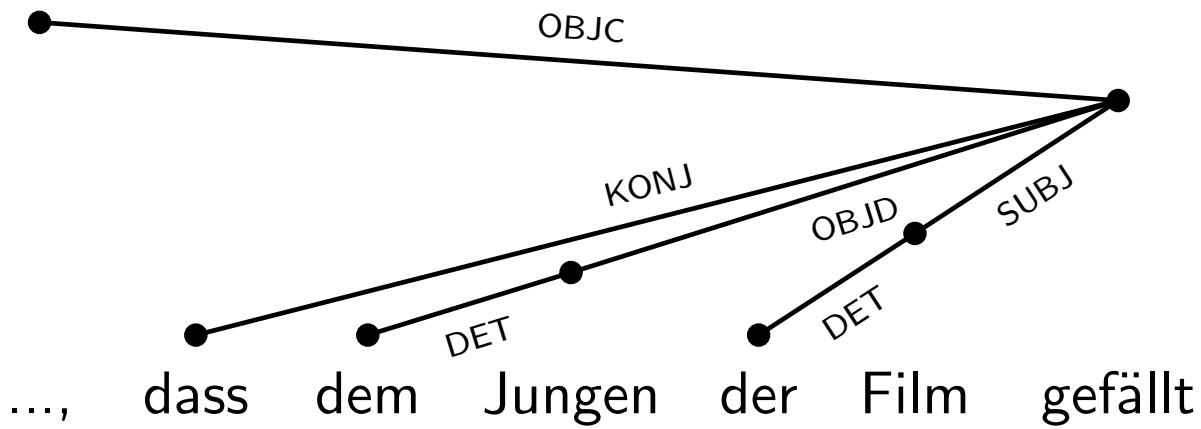
Predictive parsing

- naive incremental processing attaches a word as soon as possible
  - $\rightarrow$ eager processing, but not eager enough
  - attachment is pending until the attachment point becomes available
  - results in a fragmented (less informative) interpretation
- predictive parsing collects pending attachments using virtual nodes
  - virtual nodes can be integrated into the current structure
  - allow to build a connected (more informative) structure



Predictive parsing

- naive incremental processing attaches a word as soon as possible
  - $\rightarrow$ eager processing, but not eager enough
  - attachment is pending until the attachment point becomes available
  - results in a fragmented (less informative) interpretation
- predictive parsing collects pending attachments using virtual nodes
  - virtual nodes can be integrated into the current structure
  - allow to build a connected (more informative) structure



- partially underspecified semantic structures can be immediately derived

Predictive parsing

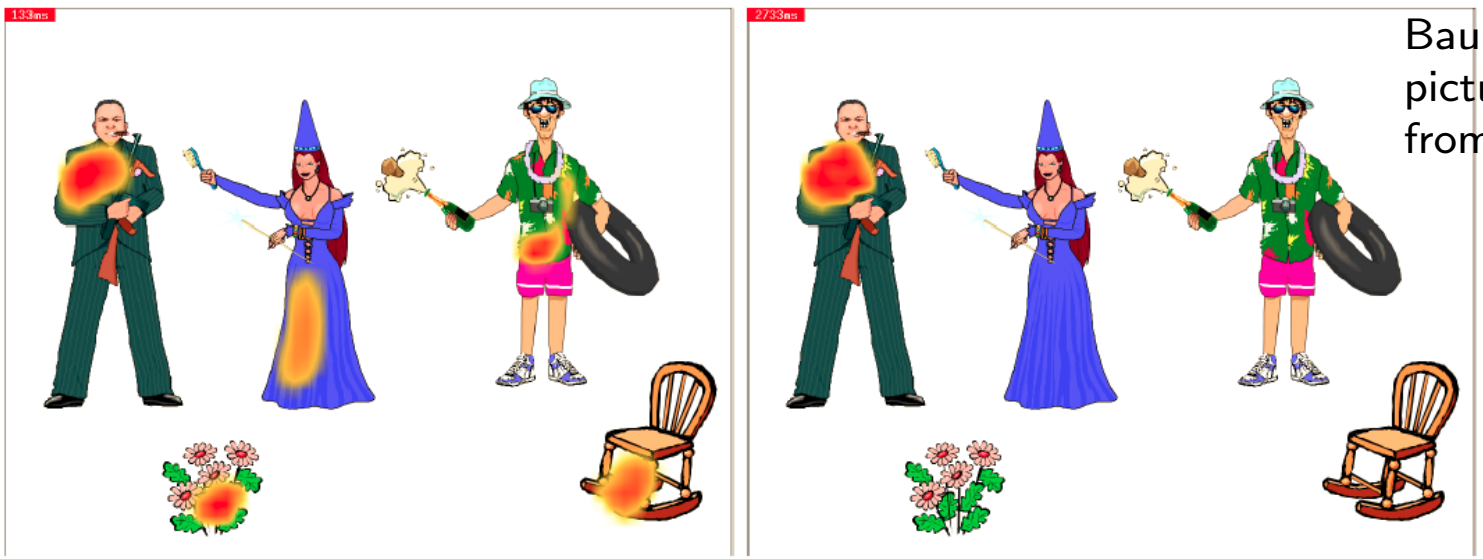
- the optimal structure is determined by the general transformation-based approach
  - reusing the result of the left context as starting point for the next increment
  - no dedicated procedural components
  - reanalysis behavior results from a shift of the optimum

Parsing in the visual world paradigm

- testing the predictive power of the parser in a similar setting (Baumgärtner 2012)
  - interfacing the parser with a bottom up model of visual attention (Itti and Koch 2001)
  - modulating the saliency landscape by means of top down information from the parser

Parsing in the visual world paradigm

- visual stimuli: three persons, two actions
  - a visually ambiguous character: the fairy
    - brushing (the gangster) and being splashed (by the tourist)
  - two visually unambiguous characters:
    - the gangster (being brushed) and the tourist (splashing)
- linguistic stimuli: ⟨ambiguous char.⟩ ⟨action⟩ ⟨unambiguous char.⟩

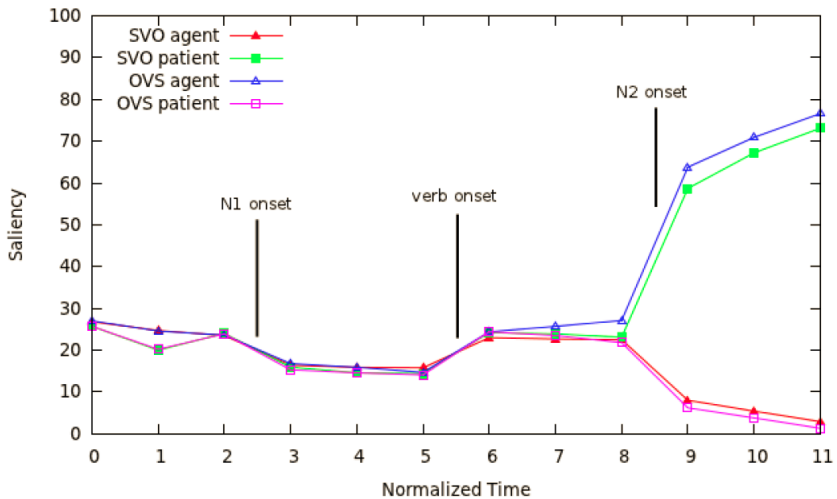


Baumgärtner (2012)
pictures adapted
from Knöferle (2005)

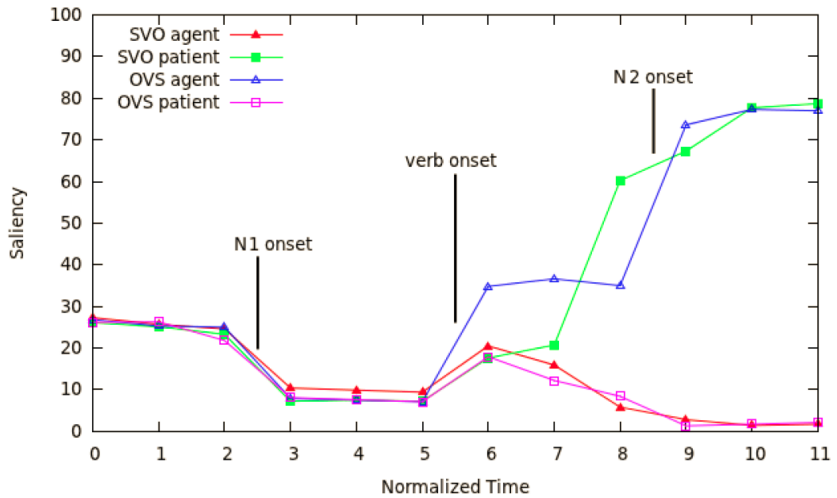
Die Fee bürstet hier den Gangster / The fairy$_{SUBJ/OBJA}$ brushes here the gangster$_{OBJA}$

Parsing in the visual paradigm

- distribution of saliency for the visual agent and the visual patient



without prediction

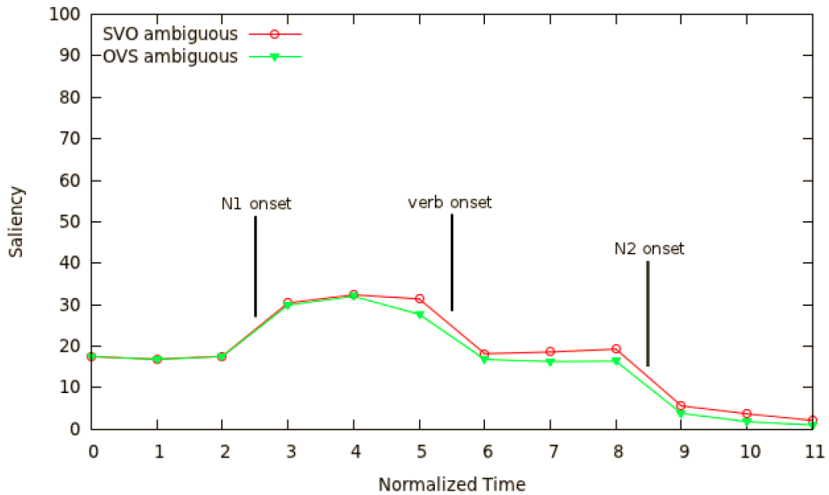


with prediction

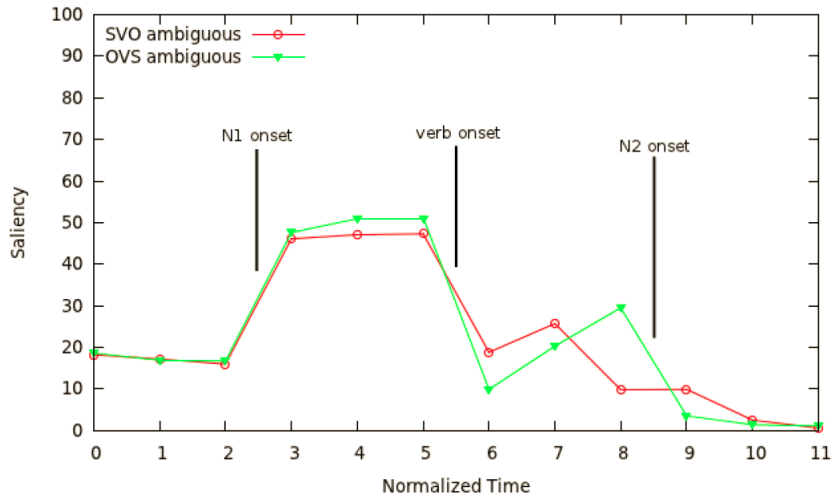
- predictions based on the available sentence prefix speed up reference resolution

Parsing in the visual world paradigm

- distribution of saliency for the ambiguous characters



without prediction

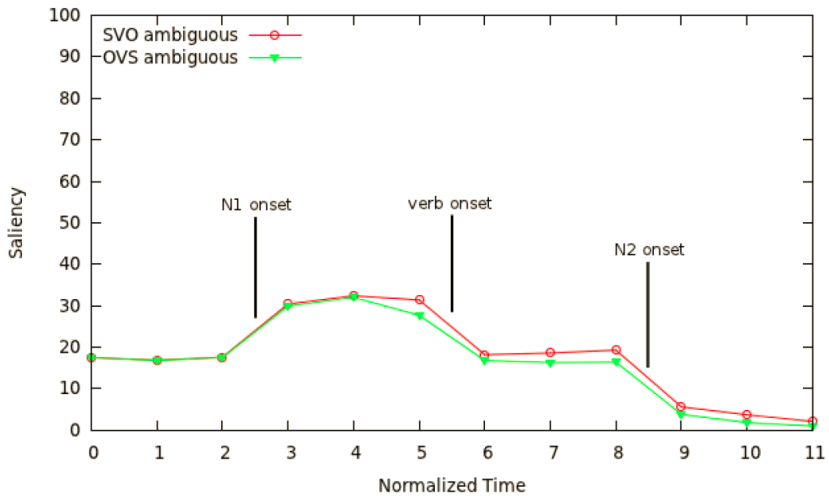


with prediction

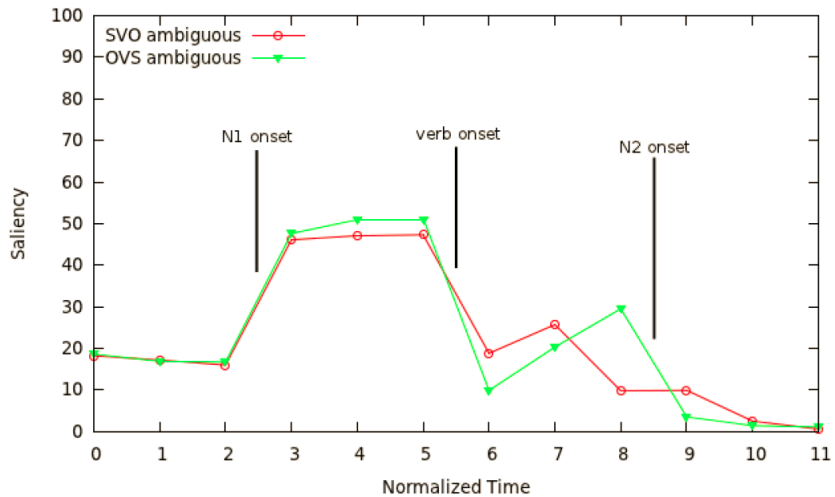
- at the first NP: higher saliency of the ambiguous character
  - not only mentioned but also *predicted* to be the subject

Parsing in the visual world paradigm

- distribution of saliency for the ambiguous characters



without prediction

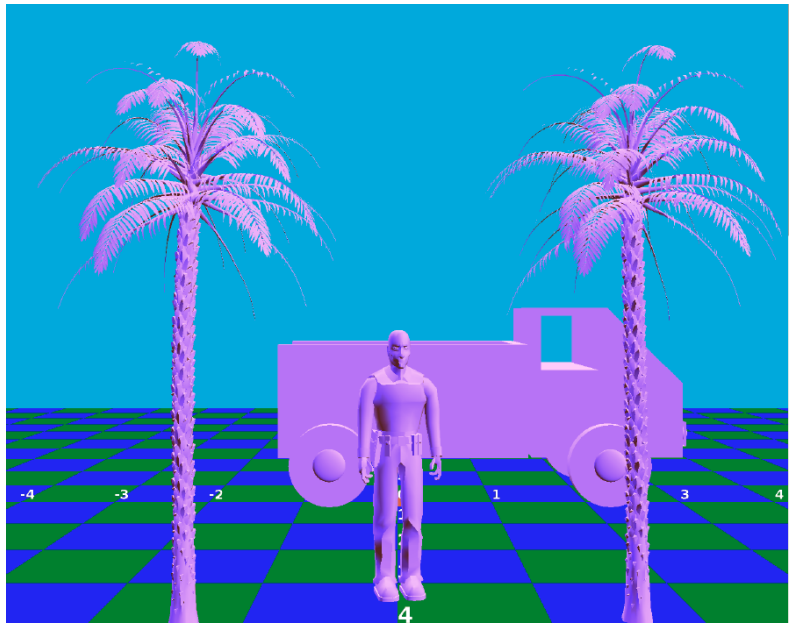


with prediction

- at the first NP: higher saliency of the ambiguous character
  - not only mentioned but also *predicted* to be the subject
- prediction produces saliency results which closely resemble findings on the human model as reported in Knöferle et al. (2005)

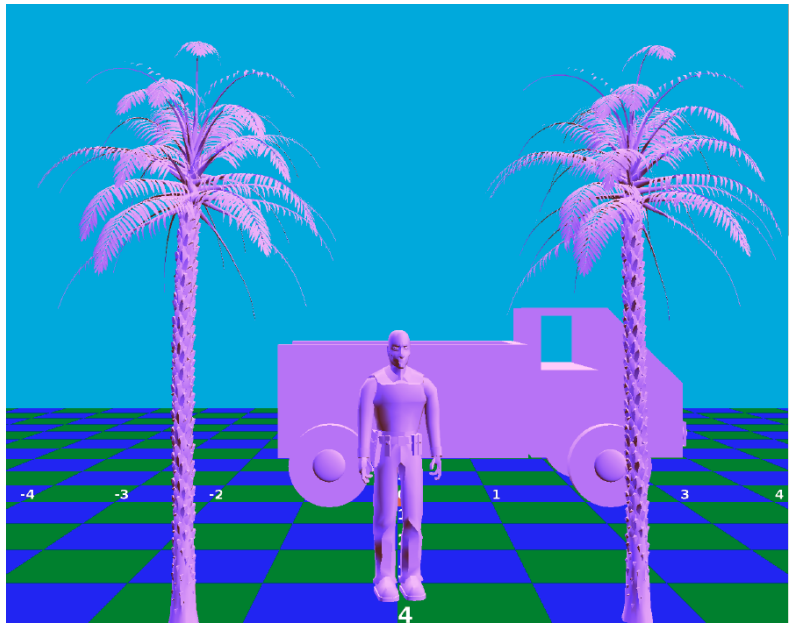
Parsing in the visual world paradigm

- experiments can also be extended to dynamic visual contexts



Parsing in the visual world paradigm

- experiments can also be extended to dynamic visual contexts



- If the visual information becomes available at different points in time. how does the parsing behaviour change?
- What's the latest point in time at which the parsing behaviour still can be affected?
- How does a changing viewpoint in the virtual world affect the parsing behaviour?

What's next?

- WCDG has nice procedural properties
  - anytime: decisive and interruptable
  - robust: never breaks down
  - hybrid: external predictors may contribute shallow syntactic or contextual cues
  - multilevel: mapping between different abstract representations
  - non-monotonic: default reasoning for efficient decision taking
  - diagnostic: the parser itself can explain the deficiencies of the current analysis

What's next?

- WCDG has nice procedural properties
  - anytime: decisive and interruptable
  - robust: never breaks down
  - hybrid: external predictors may contribute shallow syntactic or contextual cues
  - multilevel: mapping between different abstract representations
  - non-monotonic: default reasoning for efficient decision taking
  - diagnostic: the parser itself can explain the deficiencies of the current analysis
- ... but also comes with severe drawbacks
  - extremely slow (8 vs. 0.015 seconds per word)
  - requires labor-intensive manual grammar development ($\geq$ 5 PY)
  - no longer state-of-the-art in parsing quality

What's next?

- moving from hybrid parsing to fully trained models
  - TurboParser (Martins et al. 2009) based on Integer Linear Programming
  - RBG-Parser (Zhang et al. 2014) separating generation of dependency trees from scoring them
  - much more efficient, more expressive models (can capture higher order constraints)
- but lack most of the desired properties

What's next?

- Can these approaches be modified to regain them?

What's next?

- Can these approaches be modified to regain them?
- predictive parsing with Turbo-Parser (Köhn and Menzel 2014)
  - providing an additional node (UNUSED) to catch all the currently unattached virtual nodes

What's next?

- Can these approaches be modified to regain them?
- predictive parsing with Turbo-Parser (Köhn and Menzel 2014)
  - providing an additional node (UNUSED) to catch all the currently unattached virtual nodes
  - modelling the rules for dealing with virtual nodes by means of integer linear equations
    - a virtual node attached to UNUSED cannot have daughters
    - a virtual node cannot be the root without having daughters
    - only virtual nodes may be attached to UNUSED

What's next?

- Can these approaches be modified to regain them?
- predictive parsing with Turbo-Parser (Köhn and Menzel 2014)
  - providing an additional node (UNUSED) to catch all the currently unattached virtual nodes
  - modelling the rules for dealing with virtual nodes by means of integer linear equations
    - a virtual node attached to UNUSED cannot have daughters
    - a virtual node cannot be the root without having daughters
    - only virtual nodes may be attached to UNUSED
  - current limitations:
    - only pseudo incremental parsing: no reuse of partial results
    - no full reanalysis capability: edge labeling precedes attachment decisions $\rightarrow$ a label alone will never change

What's next?

- parsing into multi-level representations
  - early experiments by Amr Rekaby Salamaa

What's next?

- parsing into multi-level representations
  - early experiments by Amr Rekaby Salamaa
- studying the predictive capabilities with more varied utterances and richer visual stimuli

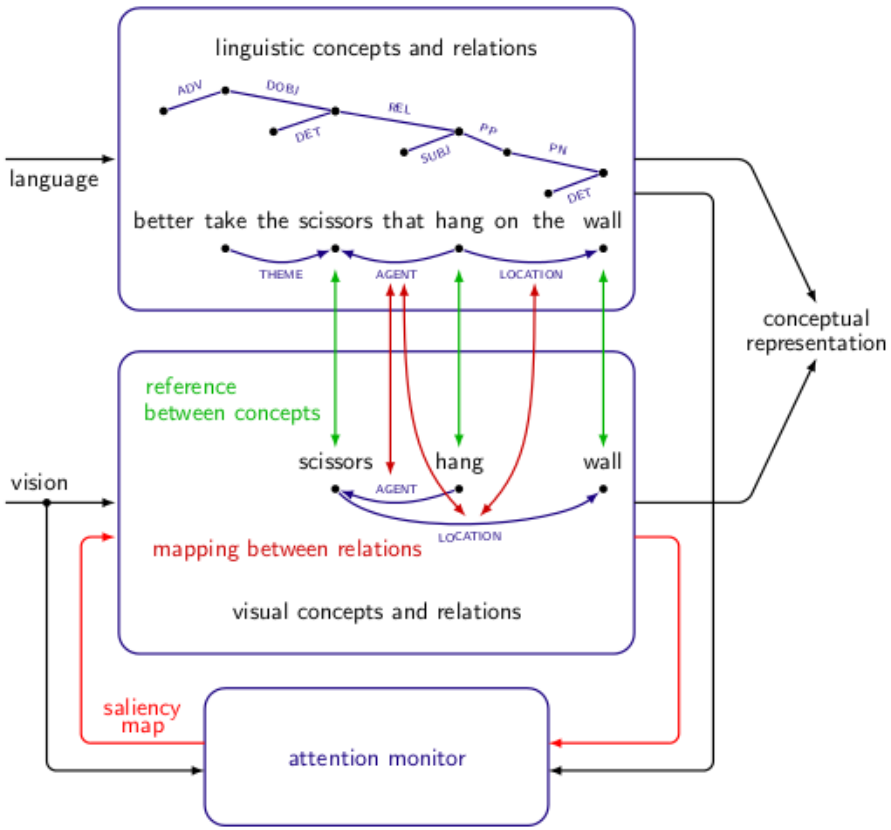
What's next?

- parsing into multi-level representations
  - early experiments by Amr Rekaby Salamaa
- studying the predictive capabilities with more varied utterances and richer visual stimuli

- exploiting the diagnostic capabilities of WCDG for language learning
  - using the visual context to determine the intended interpretation
  - research by Christine Köhn

What's next?

- experiments with active vision approaches
  - visual attention drives visual information acquisition
  - feeding the acquired information back into the visual channel
  - using visual attention as a means to acquire more detailed visual information
  - which might help to improve language comprehension



Summary

Summary

- the combination of a relational model, with weighted constraints and a transformation-based search allowed us to build an incremental, decisive, predictive, context aware, ... parser
- interfaced with a (simulated) component for visual information extraction the parser has shown a high degree of psycholinguistic adequacy
- Can such a behaviour also been achieved with machine learning techniques?