

Introduction to Corpus Resources, Annotation and Access: *Semantic Annotation*

Sabine Schulte im Walde
Universität Stuttgart

Foundational Course
Departament de Traducció i Filologia
Universitat Pompeu Fabra
April 16-20, 2007

Overview

1. Word senses and Word Sense Disambiguation
2. Lexical semantic resources
3. Sense and role annotation and evaluation

Word Senses and Word Sense Disambiguation

Word Senses

- **Lexical semantics** is the study of how and what the words of a language denote.
- Lexical semantics involves the meaning of each individual word.
- A **word sense** is one of the meanings of a word.
- A word is called **ambiguous** if it can be interpreted in more than one way, i.e., if it has multiple senses.
- **Disambiguation** determines a specific sense of an ambiguous word.

Homonymy and Polysemy

- A **homonym** is a word with multiple, unrelated meanings. A homonym is a word that is spelled and pronounced the same as another but with a different meaning.
 - bank* → financial institution
 - slope of land alongside a river
- A **polyseme** is a word with multiple, related meanings.
 - school* → *He goes to school every day.* (institution)
 - *The school has a blue facade.* (building)
 - *The school is on strike.* (teacher)
- **Regular polysemy** performs a regular induction of a word sense on the basis of another, cf. *school / office*.

Human Beings and Ambiguity

- What seems perfectly obvious to a human being is deeply ambiguous to the computer, and there is no easy way of resolving ambiguity.
 - » *I paid the money on my **bank** account.*
 - » *I watched the ducks on the river **bank**.*
- Semantic priming (psycholinguistics):
The response time for a word is reduced when it is presented with a semantically related word.
 - doctor** → nurse / butter*
- If an ambiguous prime such as *bank* is given, it turns out that all word senses are primed for.
 - bank** → money / river*

Disambiguation Cues

- Probability and prototypicality → default interpretation:
corpus-related importance of word senses
- Internal text evidence (context), in particular collocations:
words, morpho-syntactic embedding, etc.
- One sense per discourse
- Domain
- Real-world knowledge

Word Sense Disambiguation (WSD)

- WSD involves the association of a given word in a text or discourse with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to that word.
- Intermediate task: not an end in itself, but necessary in most NLP tasks, such as machine translation, information retrieval, speech processing
- Problems:
 1. Which are the senses?
 2. Which is the correct sense?
- Major sources:
 1. Context of the word to be disambiguated
 2. External knowledge sources

Sense Inventory

- Word Sense Disambiguation needs a set of word senses to disambiguate between.
- Sense inventories are found in dictionaries, thesauri or similar.
- The granularity and criteriae for the set of senses differ.
- There is no reason to expect a single set of word senses to be appropriate for different NLP applications.

Word Senses: References

- Michael Lesk (1986): “Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone”. In *Proceedings of the SIGDOC Conference*. Toronto, Canada.
- William A. Gale, Kenneth W. Church, and David Yarowsky (1992): “One sense per discourse”. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 233-237.
- Adam Kilgarriff (1997): “I don't believe in word senses”. *Computers and the Humanities* 31(2):91-113.
- Patrick Hanks (2000): “Do word meanings exist?” *Computers and the Humanities*, 34(1-2):205-215.
- Martha Palmer (2000): “Consistent criteria for sense distinctions”. *Computers and the Humanities*, 34(1-2):217-222.

Lexical Semantic Resource



Lexical Semantic Resources

Sense inventory and organisation:

- WordNet

Sense annotation and semantic role annotation:

- Prague Dependency Treebank
- FrameNet
- PropBank
- OntoBank / OntoNotes

WordNet

- Online lexical reference system
- The design is inspired by current psycholinguistic theories of **human lexical memory**.
- English nouns, verbs, adjectives and adverbs are organised into **synonym sets** (synsets).
- Each synset represents one underlying **lexical concept**.
- Different **(paradigmatic) relations** link the synonym sets.
- WordNet was developed by the Cognitive Science Laboratory at Princeton University under the direction of George A. Miller.
- WordNets now exist for many languages.

WordNet Synsets

- Synsets are sets of synonymous words.
- Polysemous words appear in multiple synsets.
- Examples:

{coffee, java}

noun example

{coffee, coffee tree}

{coffee bean, coffee berry, coffee}

{chocolate, coffee, deep brown, umber, burnt umber}

{cold}

adjective example

{aloof, cold}

{cold, dry, uncordial}

{cold, unaffectionate, uncaring}

{cold, old}

WordNet Relations

Within synsets:

- **Synonymy**, such as {coffee, java}

Between synsets / parts of synsets:

- **Antonymy**: opposition, such as {cold} — {hot}
- **Hypernymy / Hyponymy**: is-a relation, such as {coffee, java} — {beverage, drink, potable}
- **Meronymy / Holonymy**: part-of relation, such as {coffee bean, coffee berry, coffee} — {coffee, coffee tree}

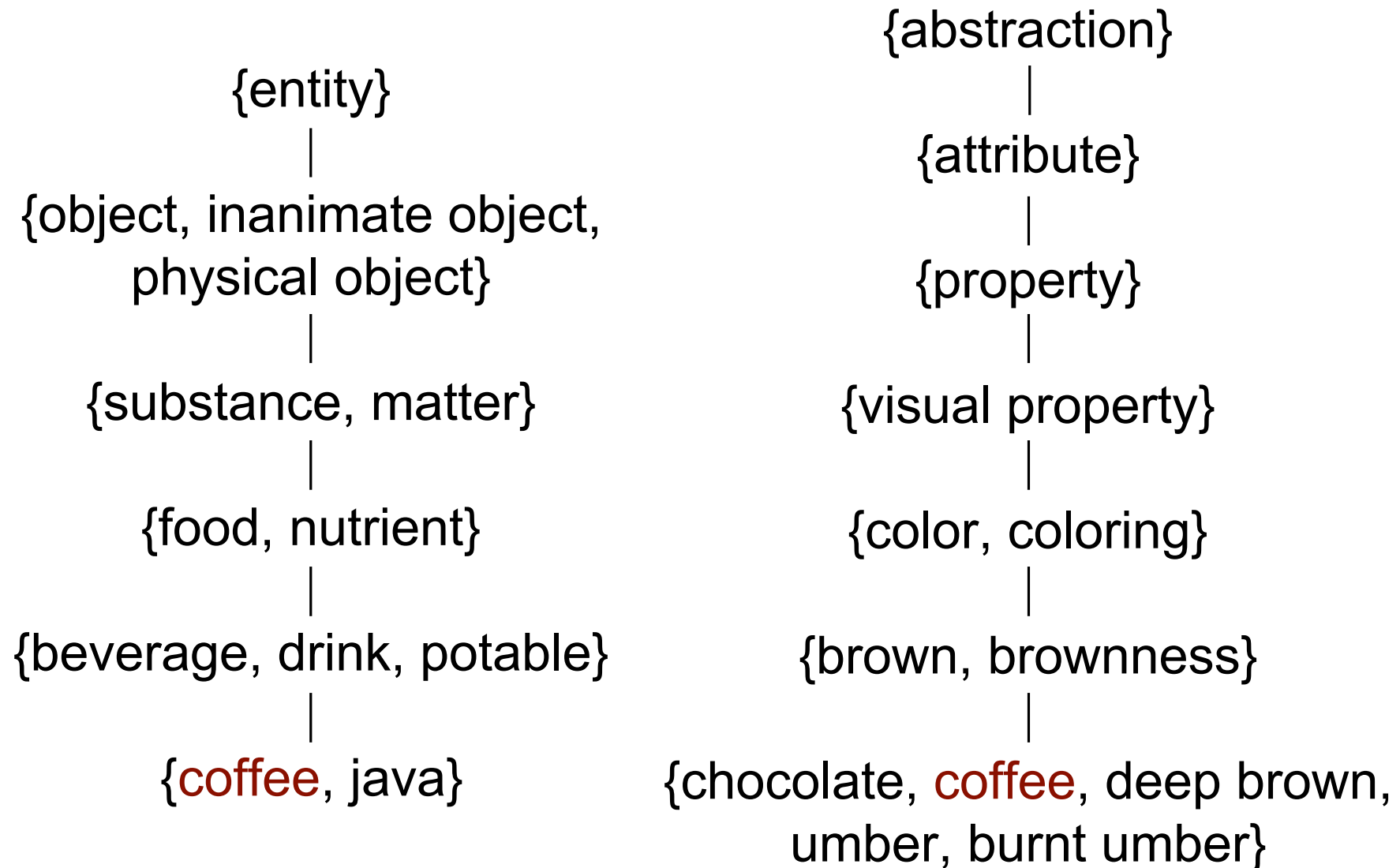
Morphology:

- **Compounds**: arabian coffee, coffee break, coffee table

WordNet Hierarchy

- Depending on the part-of-speech, different relations are defined for a word. For example, the core relation for nouns is hypernymy, the core relation for adjectives is antonymy.
- Hypernymy imposes a hierarchical structure on the synsets.
- The most general synsets in the hierarchy consists of a number of pre-defined disjunctive top-level synsets:
 - nouns → {entity}, {abstraction}, {psychological}, etc.
 - verbs → {move}, {change}, {get}, {feel}, etc.

WordNet Hierarchy: Examples



WordNet Family

- Current status: WordNets for 38 languages
- WordNets in the world:
http://www.globalwordnet.org/gwa/wordnet_table.htm
- Integration of WordNets into multi-lingual resources:
 - » EuroWordNet: English, Dutch, Italian, Spanish, German, French, Czech and Estonian
 - » BalkaNet: Bulgarian, Czech, Greek, Romanian, Turkish, Serbian
- An inter-lingual index connects the synsets of the WordNets.

WordNet: References

- George A. Miller, editor (1990): "WordNet: An on-line lexical database". Special issue of the *International Journal of Lexicography*, 3(4).
- Christiane Fellbaum, editor (1998): "WordNet - An electronic lexical database". MIT Press.
- Piek Vossen (2004): "EuroWordNet: A multilingual database of autonomous and language-specific WordNets connected via an interlingual index". *International Journal of Lexicography*, 17:161-173.
- WordNets online:
 - » WordNet: <http://wordnet.princeton.edu/>
 - » EuroWordNet: <http://www.ilc.uva.nl/EuroWordNet/>
 - » MultiWordNet: <http://multiwordnet.itc.it/>
 - » Global WordNet Association: <http://www.globalwordnet.org/>

Excursus: Semantic Roles and Verb Alternations

Classical Thematic Roles (1)

- **Agent**: animate, volitional; initiates action
Anna prepared chicken for dinner.
- **Patient**: animate or inanimate; undergoes (and is affected by) action
Anna baked a cake for her daughter.
- **Experiencer**: animate; undergoes perceptual experience
The storm frightened Anna.
- **Theme**: animate or inanimate; undergoes motion, or an action that does not affect it significantly
Anna sent Tim a letter.
- **Recipient**: generally animate; receives something
Tim kicked Bob the ball.

Classical Thematic Roles (2)

- **Benefactive**: generally animate; one who benefits from the event
Anna baked a cake for her daughter.
- **Goal**: animate or inanimate; endpoint of the action
Anna put the book on the table.
- **Location**: place where the event occurs
Anna and Tim met in Paris.
- **Source**: animate or inanimate; starting point of an action
Anna and Tim came from Berlin.
- **Instrument**: often inanimate; used in an action
Tim smashed the window with a hammer.

Diathesis Alternations (examples)

- Unergative:
 - The horse raced past the barn.
 - The jockey raced the horse past the barn.
- Unaccusative:
 - The butter melted in the pan.
 - The cook melted the butter in the pan.
- Object-Drop:
 - The boy played.
 - The boy played soccer.

[Examples taken from Paola Merlo and Suzanne Stevenson (2001):
“Automatic verb classification based on statistical distributions of
argument structure”. *Computational Linguistics*, 27(3):373-408.]

Prague Dependency Treebank (PDT)

- Three-level annotation scenario:
 1. **morphological level**
 2. syntactic annotation at the **analytical level**
 3. linguistic meaning at the **tectogrammatical level**
- Corpus data:
 - newspaper articles (60%), economic news and analyses (20%), popular science magazines (20%)
- 1.8 million tokens are annotated on the morphological level, 1.0 million tokens are annotated on the tectogrammatical level.

Tectogrammatical Level of the PDT

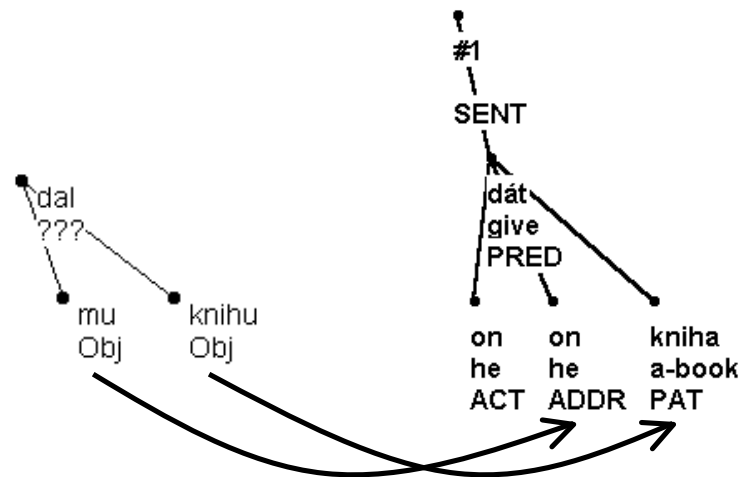
- Annotation: dependency, functor, grammemes, ellipsis solution, coreference, topic/focus (deep word order), valency lexicon
- 39 attributes
- Similar to the surface (analytical) level ... but:
 - » certain nodes deleted
(auxiliaries, non-autosemantic words, punctuation)
 - » some nodes added
(based on word - mostly verb, noun - valency)
 - » some ellipsis resolution
(detailed dependency relation labels: functors)

Tectogrammatical Functors

- General functors, e.g.:
actor/bearer, addressee, patient, origin, effect, cause, regard, concession, aim, manner, extent, substitution, accompaniment, locative, means, temporal, attitude, cause, regard, directional, benefactive, comparison
- Specific functors for dependents on nouns, e.g.:
material, appurtenance, restrictive, descriptive, identity
- Subtle differentiation of syntactic relations, e.g.:
temporal (before, after, on),
accompaniment, regard, benefactive (for/against)

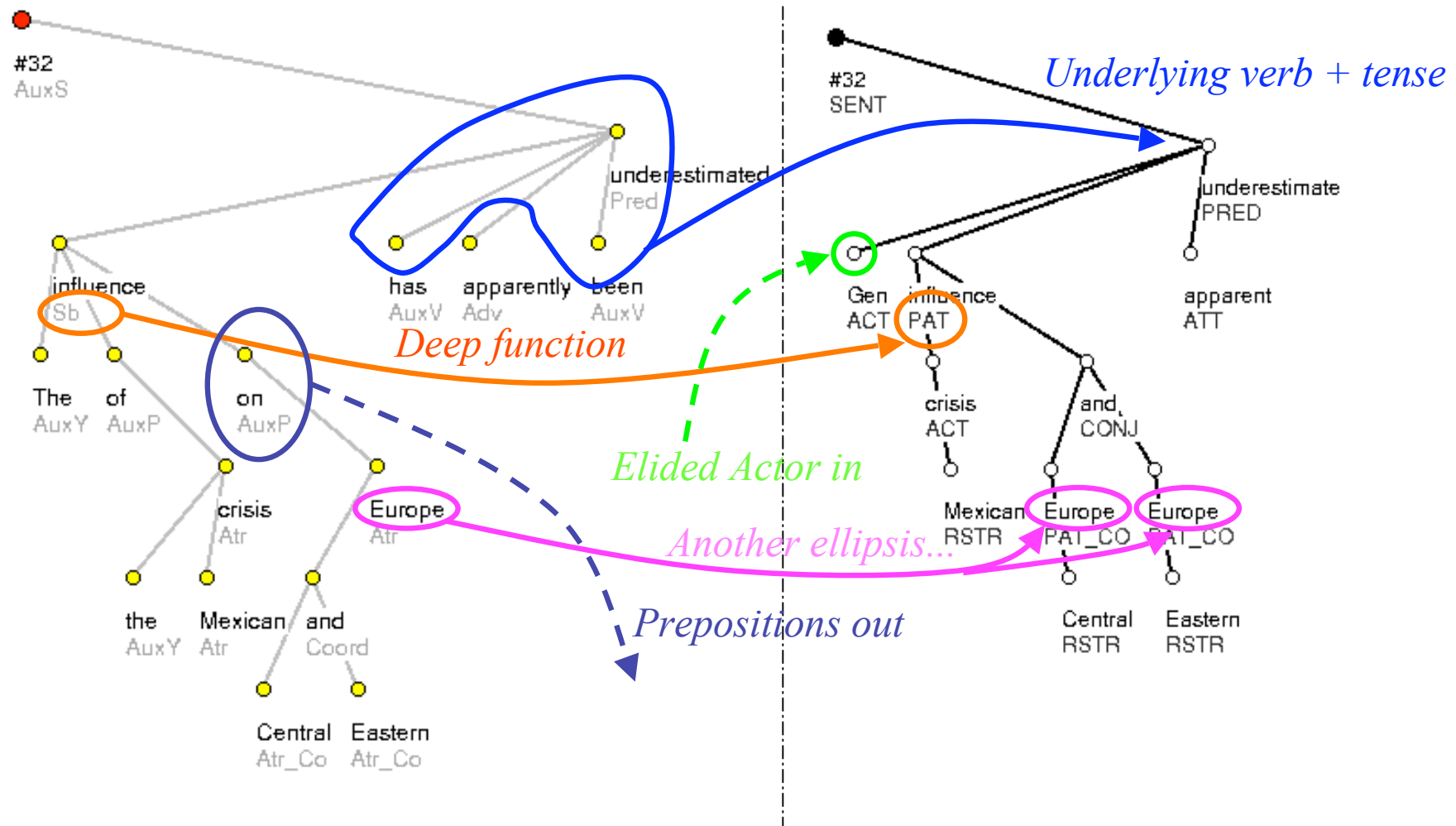
Tectogrammatical Example

- Example: *(he) gave him a book*
dal mu knihu



The “Obj” goes into ACT, PAT, ADDR, EFF or ORIG,
as based on the governor’s valency frame.

Analytical vs. Tectogrammatical Level



PDT: References

- Eva Hajičová, Jarmila Panevová, and Petr Sgall (2000): "A manual for tectogrammatic tagging of the Prague Dependency Treebank". UFAL/CKL Technical Report TR-2000-09, Charles University, Prague.
- Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká (2003): "The Prague Dependency Treebank: A three-level annotation scenario". In: Anne Abeille, editor: "Treebanks: building and using syntactically annotated corpora". Kluwer Academic Publishers.
- Petr Sgall, Jarmila Panevová, and Eva Hajičová (2004): "Deep syntactic annotation: Tectogrammatical representation and beyond". In *Proceedings of the HLT-NAACL Workshop on "Frontiers in Corpus Annotation"*. Boston, MA.
- Jan Hajič and Zdeňka Urešová (2005): "The Prague Dependency Treebank and Valency Annotation". Tutorial at RANLP, Borovets.
- PDT online: <http://ufal.mff.cuni.cz/pdt/>

FrameNet

- Frame-semantic descriptions for English verbs, nouns, and adjectives
- Aim: document the range of **semantic and syntactic combinatory possibilities (valences)** of each word in each of its senses
- Result: lexical database with
 - » descriptions of the **semantic frames**
 - » a **representation of the valences** for target words
 - » a collection of **annotated corpus attestations**
- Current size: more than 6,100 lexical units annotated in more than 625 semantic frames, exemplified in more than 135,000 sentences

FrameNet Vocabulary

- **Frame semantics**, developed by Charles Fillmore:
 - » a theory that relates linguistic semantics to encyclopaedic knowledge
 - » describes the meaning of a word (sense) by characterising the essential background knowledge that is necessary to understand the word/sentence
- **Frame**: conceptual structure modelling prototypical situations
- **Frame element**: frame-evoking word or expression
- **Frame roles**: participants and properties of the situation

FrameNet Example

- **Frame:** *Transportation*
 - » **Frame elements:** mover, means, path
 - » **Scene:** mover moves along path by means
- **Frame:** *Driving*
 - » Inherit: Transportation
 - » **Frame elements:** driver=mover, rider=mover, cargo=mover, vehicle=means
 - » **Scenes:** driver starts vehicle, driver controls vehicle, driver stops vehicle
- Annotated corpus sentence:
Now [_D Tim] was driving [_R his guest] [_P to the station].

FrameNet Languages

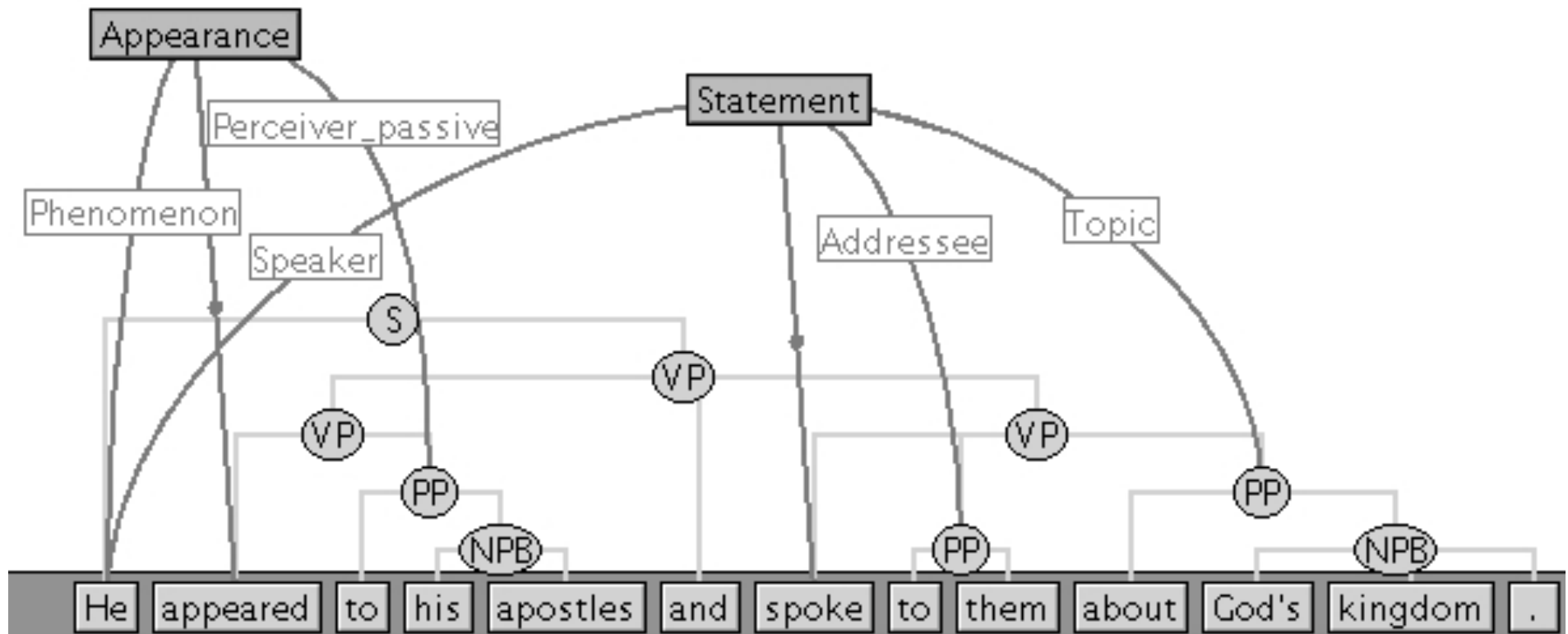
- English FrameNet: Berkeley
- German FrameNet: Salsa, Saarbrücken
- Spanish FrameNet: Barcelona
- Japanese FrameNet: Keio, Yokohama & Tokyo

Issue: cross-lingual transfer of English FrameNet

German FrameNet: SALSA

- Annotation of the **TIGER treebank** with **semantic roles**
- Existing manual syntactic annotation of newspaper data: grammatical functions, syntactic categories, argument structure of syntactic heads
- Annotation procedure: All frame elements are annotated by their frames and roles → **corpus-based**.
(In comparison: The English FrameNet annotates a selected set of prototypical examples for each frame → frame-based.)
- Current size: 476 German predicates with 18,500 instances and 628 different frames

TIGER/SALSA Example



FrameNet: References (1)

[General & English:]

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe (1998): "[The Berkeley FrameNet project](#)". In *Proceedings of the 17th International Conference on Computational Linguistics*, pp. 86-90.
- Thierry Fontenelle, editor (2003): "[FrameNet and frame semantics](#)". Special issue of the *International Journal of Lexicography*, 16(3).

[German:]

- Katrin Erk, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal (2003): "[Towards a resource for lexical semantics: a large German corpus with extensive semantic annotation](#)". In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.
- Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, Manfred Pinkal (2006): "[The SALSA corpus: a German corpus resource for lexical semantics](#)". In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

FrameNet: References (2)

[Spanish:]

- Carlos Subirats and Hiroaki Sato (2004): "[Spanish FrameNet and FrameSQL](#)". In *Proceedings of the LREC Workshop on "Building Lexical Resources from Semantically Annotated Corpora"*.

[Japanese:]

- Kyoko Hirose Ohara, Seiko Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito, and Shun Ishizaki (2004): "[The Japanese FrameNet project: An introduction](#)". In *Proceedings of the LREC Workshop on "Building Lexical Resources from Semantically Annotated Corpora"*.

FrameNet online:

- » English: <http://framenet.icsi.berkeley.edu/>
- » German: <http://www.coli.uni-saarland.de/projects/salsa/>
- » Spanish: <http://gemini.uab.es/SFN/>
- » Japanese: <http://jfn.st.hc.keio.ac.jp/>

PropBank (1)

- The PropBank project creates a corpus of text annotated with information about basic semantic propositions.
- A layer of **predicate-argument relations (semantic roles)** is added to the syntactic trees of the **Penn Treebank**.
- The Penn Treebank does not distinguish the different roles played by a verb's grammatical functions. Because the same verb used with the same syntactic sub-categorisation can assign different semantic roles, roles cannot be deterministically added to the Treebank by an automatic process.
- Every instance of every verb in the corpus is covered.

PropBank (2)

- Goal: broad-coverage hand-annotated corpus of semantic roles → **verb alternations**
- Basis: linking between semantic roles and syntactic realisation; the syntactic frames are a direct reflection of the underlying semantics (Levin, 1993)
- Purpose: useful level of semantic representation and a corpus of annotated data to enable empirical studies
- Potential applications: information extraction, question answering, machine translations

PropBank Procedure

- **Framing:** collection of framesets for each lexeme
 1. examine a sample of corpus sentences for the verb under consideration
 2. group the instances into one or more major senses
 3. turn each major sense into a single frameset
- **Annotation:**
 1. run a rule-based argument tagger on the corpus; 83% accuracy on pilot data
 2. correct the tagger output by hand, on a verb by verb basis
 3. adjudicate to resolve differences between annotators

PropBank Roles

- PropBank defines **semantic roles on a verb by verb basis**.
- An individual verb's semantic arguments are numbered, beginning with 0. For a particular verb, *Arg0* is generally the argument exhibiting features of a prototypical **agent** while *Arg1* is a prototypical **patient** or **theme**.
- **Roleset**: set of roles for to a distinct usage of a verb
- **Frameset**: roleset associated with a set of syntactic frames indicating allowable syntactic variations
- The numbered arguments plot a middle course among many different theoretical viewpoints and can be mapped onto any theory of argument structure.

PropBank *ArgM* Modifier Roles

PropBank also defines **roles that can apply to any verb:**

- LOC: location
- EXT: extent
- DIS: discourse connectives
- ADV: general-purpose
- NEG: negation marker
- MOD: modal verb
- CAU: cause
- TMP: time
- PCN: purpose
- MNR: manner
- DIR: direction

PropBank Examples

- Frameset *accept*₁ “take willingly”

Arg0: acceptor

Arg1: thing accepted

Arg2: accepted-from

Arg3: attribute

[Arg0 He] [ArgM-mod would] [ArgM-neg n't] accept [Arg1 anything of value] [Arg2 from those he was writing about].

- Frameset *kick*₁ “drive or impel with the foot”

Arg0: kicker

Arg1: thing kicked

Arg2: instrument (defaults to foot)

*[Arg0 John_i] tried [Arg0 *trace*_i] to kick [Arg1 the football].*

PropBank Polysemy (1)

- A polysemous verb may have more than one frameset.
- Frameset *decline*₁ “go down incrementally”
 - Arg1: entity going down
 - Arg2: amount gone down by EXT
 - Arg3: start point
 - Arg4: end point

... [*Arg*₁ *its income*] declining [*Arg*_{2-EXT} 42%] [*Arg*₄ *to \$2,420*].
- Frameset *decline*₂ “demure, reject”
 - Arg0: agent
 - Arg1: rejected thing

[*Arg*₀ *A spokesman*_{*i*}] declined [*Arg*₁ **trace**_{*i*} *to elaborate*].

PropBank Polysemy (2)

- Alternations which preserve verb meanings, such as causative/inchoative or object deletion, are considered to be one frameset only.

- Frameset *open*₁ “cause to open”

Arg0: agent

Arg1: thing opened

Arg2: instrument

[Arg0 John] opened [Arg1 the door].

[Arg1 The door] opened.

[Arg0 John] opened [Arg1 the door] [Arg2 with his foot].

PropBank vs. (English) FrameNet

- Common goal: document the syntactic realisation of arguments of the predicates of the general English lexicon by annotating a corpus with semantic roles
- Different methodologies:
 - » FrameNet proceeds on a frame-by-frame basis.
 - » PropBank is aimed at providing data for training statistical systems and has to provide an annotation for every clause in the Penn Treebank.
 - » PropBank puts less emphasis on the definition of the semantics of the class the verbs are associated with.
 - » The PropBank semantic roles may not correspond to the FrameNet frame semantic elements.

PropBank: References

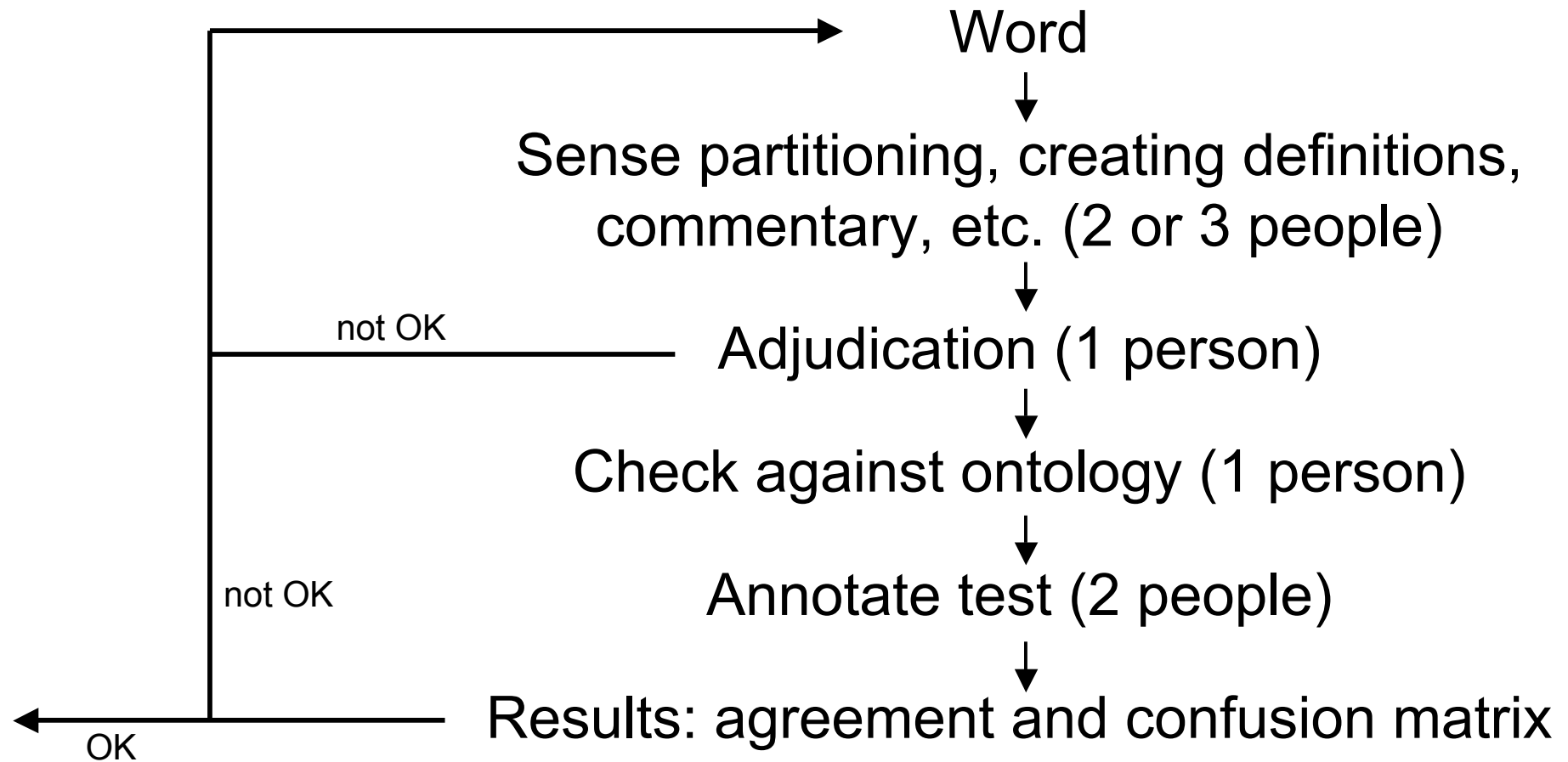
- Paul Kingsbury and Martha Palmer (2002): “From Treebank to PropBank”. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Spain.
- Martha Palmer, Dan Gildea, and Paul Kingsbury (2005): "The Proposition Bank: A corpus annotated with semantic roles". *Computational Linguistics*, 31(1):71-106.
- Olga Babko-Malaya (2005): “PropBank annotation guidelines”. MS.
- Olga Babko-Malaya (2005): “Guidelines for PropBank framers”. MS.
- PropBank online:
<http://verbs.colorado.edu/mpalmer/palmer/projects/ace.html>
- NomBank online:
<http://nlp.cs.nyu.edu/meyers/NomBank.html>

OntoBank / OntoNotes

- **Goal:** domain-independent representation of literal meaning that includes predicate structure, word sense, ontology linking, and coreference
- **Languages:** English, Chinese, Arabic
- **Genres:** newswire, news groups, weblogs, etc.
- Sense distinctions are represented by linguists in a hierarchical structure, similar to a decision tree.
- ISI's Omega Ontology: <http://omega.isi.edu/>
- **Partners:** BBN (Weischedel), University of Colorado (Palmer), University of Pennsylvania (Marcus), ISI (Hovy)

OntoBank: Annotation Procedure

The 90% solution



OntoBank: Annotation Procedure

The 90% solution

The most frequent noun and verb senses in a 300K subset of the PropBank are annotated:

- A 50-sentence sample of instances is annotated and immediately checked for inter-annotator agreement.
- ITA scores below 90% lead to a revision and clarification of the groupings by the linguist.
- Only after the groupings have passed the ITA hurdle, each individual group is linked to a conceptual node in the ontology.

In addition to higher accuracy, a three-fold increase in annotator productivity is found.

OntoBank: References

- Invited talk by Eduard Hovy at LREC 2006: “Corpus creation by annotation”. Genoa, Italy.
- Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel (2006): “OntoNotes: The 90% Solution”. In *Proceedings of the Human Language Technology of the North American Chapter of the Association for Computational Linguistics*. New York City, NY.
- OntoBank online: link not yet available, check Eduard Hovy's website <http://www.isi.edu/natural-language/people/hovy.html>

Automatic Sense and Role Annotation



Automatic Sense/Role Annotation

- Tasks:
 1. Definition of sense/role inventory
 2. Annotation of senses/roles
- Lexical semantic resources partly rely on (semi-) automatic sense/role annotation
- Natural Language Processing applications assume sense/role annotation, e.g. machine translation

Sense Annotation in SENSEVAL (1)

- SENSEVAL: open evaluation exercise for WSD
- SENSEVAL is the SIGLEX international organisation devoted to the evaluation of WSD systems.
- Gold standard for word sense annotation: SemCor (a.o.)
- Sense inventory taken from WordNet
- Task: develop a supervised/unsupervised system for sense annotation (all words, lexical sample)
- Subtasks for subcategorisation acquisition, etc.
- Basis: mixture of labeled and unlabeled data

Sense Annotation in SENSEVAL (2)

- Purposes:
 - » agree on an explicit and detailed definition of the task
 - » produce a gold standard corpus of correct answers
 - » evaluate the strengths and weaknesses of WSD programs with respect to different words, different varieties of language, and different languages
 - » further the understanding of lexical semantics and polysemy
- Languages: English, Italian, Basque, Catalan, Chinese, Romanian, Spanish

Shared Task on Role Labeling

- Shared task at the Conferences on Computational Language Learning (CoNLL) in 2004 and 2005
- Shallow semantic parsing: **semantic role labeling**
- Task: develop a machine learning system
 1. All the constituents in the sentence which fill a semantic role of the verb have to be recognised.
 2. Label the arguments with their semantic roles.
- Evaluation: precision, recall and F_1 measure
- The organisation provided training, development, and test sets from the standard sections of the Penn Treebank and PropBank corpora.

Sense/Role Annotation: References

- Nancy Ide and Jean Véronis (1998): “Introduction to the special issue on Word Sense Disambiguation: The state of the art”. *Computational Linguistics*, 24(1):1-40.
- Adam Kilgarriff and Martha Palmer (2000): “Introduction to the special issue on SENSEVAL”. *Computers and the Humanities*, 34(1-2):1-13.
- Philip Edmonds and Adam Kilgarriff (2002): “Introduction to the special issue on evaluating Word Sense Disambiguation systems”. *Journal of Natural Language Engineering*, 8(4).
- Xavier Carreras and Lluís Màrquez (2004,2005): “Introduction to the CoNLL shared task: semantic role labeling”. *Conferences on Natural Language Learning*.
- SENSEVAL online: <http://www.senseval.org/>