

Grammar-based treebank – a happy marriage of empiricism and theory?

Alexandr Rosen

Institute of Theoretical and Computational Linguistics
Faculty of Arts, Charles University, Prague

Grammar and Corpora 2012
4th International Conference
Czech Academy of Sciences, Prague
28–30 November 2012

The bottom line (or two)

A corpus is an approximation of language use,
a grammar is an approximation of language system.



The empirical and the theoretical sides of linguistics
meet in the annotation of a corpus.

Outline of the talk

- 1 Why treebanks, why grammars?
- 2 Treebanks
- 3 Grammars
- 4 The grammar–treebank relationship
- 5 Czech treebanking
- 6 Architecture
- 7 Examples
- 8 Input processing
- 9 Conclusions and plans

Outline of the talk

- 1 Why treebanks, why grammars?
- 2 Treebanks
- 3 Grammars
- 4 The grammar–treebank relationship
- 5 Czech treebanking
- 6 Architecture
- 7 Examples
- 8 Input processing
- 9 Conclusions and plans

Why treebanks?

Treebank ... a text corpus annotated (at least) with syntactic structure

- = why corpora?
- = why annotation?
- = why syntax?
- ?= why grammars?

Why treebanks? (cont'd)

Explicit markup of syntactic relations (constituents, heads/dependents)

→

- Easier to identify semantic relations (predicates and arguments)
- Simplifies some queries
- Simplifies extraction of lexical properties (valency)
- Support for grammar development
- Training data for NLP applications

Why grammars? 1/2

“Every time I fire a linguist, system performance goes up.”

Fred Jelinek, **1980s**

- But maybe we don't care about system performance?
- Moreover:
 - No longer a wise strategy for NLP
 - Empirical and symbolic methods can be combined
 - ‘Deep’ linguistics needed for long-term success



Why grammars? 2/2



“We should probably all spend more time on the linguistic annotation of actual data rather than on writing grammar rules, based primarily on introspection.”

Erhard Hinrichs, 1990s

- But what kind of annotation?
- “A sentence has as many structures as there are theories.”
[Haider(1993)]

Outline of the talk

- 1 Why treebanks, why grammars?
- 2 Treebanks**
- 3 Grammars
- 4 The grammar–treebank relationship
- 5 Czech treebanking
- 6 Architecture
- 7 Examples
- 8 Input processing
- 9 Conclusions and plans

Treebanks

- First treebank: *Lancaster-Leeds Treebank*
early 1980s, 45 KW, later SUSANNE, due to Geoffrey Sampson
- First major project: *Penn Treebank*
release 0.5 in 1992, now 3 MW
- Now according to Wiki: 74 treebanks in about 40 languages
- *The 11th International Workshop on Treebanks and Linguistic Theories* starts **today**, approx. 20 contributions each year

Treebanks differ in:

- Size
- Linguistic background
- Format
- Level of detail
- Depth of analysis
- Ways they are built

Also spoken, parallel, historical, ... treebanks

Treebanks around the world *)

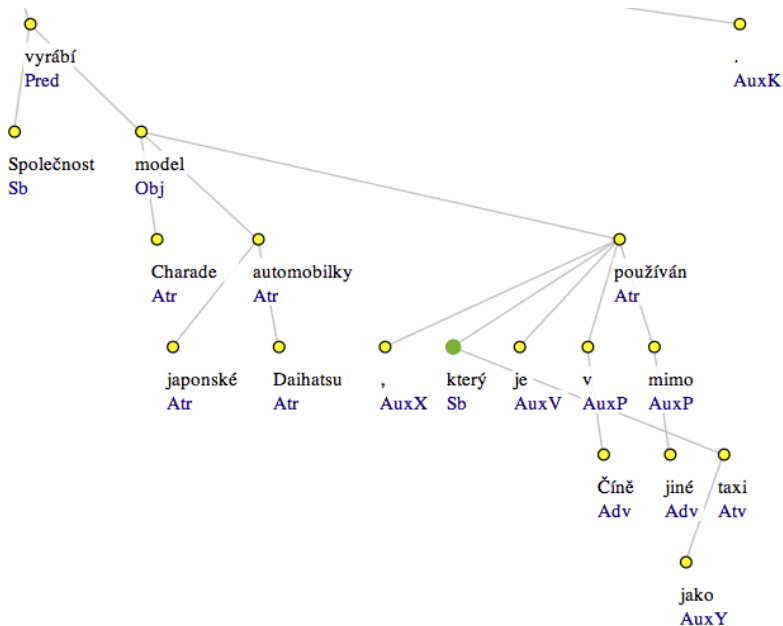
- 63 treebanks, 36 languages, sizes up to 1.5 billion words
- Also spoken (8), historical (7), parallel (4)
- Mostly stochastically parsed and manually corrected
- 15 parsed by a symbolic grammar (LFG, HPSG, DCG) and manually disambiguated
- 39 PS-based annotation, 20 dependency-based annotation
- 15 available with multiple annotation formats – *Penn Treebank*: PS, P/A, dependency, LFG, HPSG, CCG, LTAG, PDT
- 20 with on-line search interface

*) The speaker's time permitting!

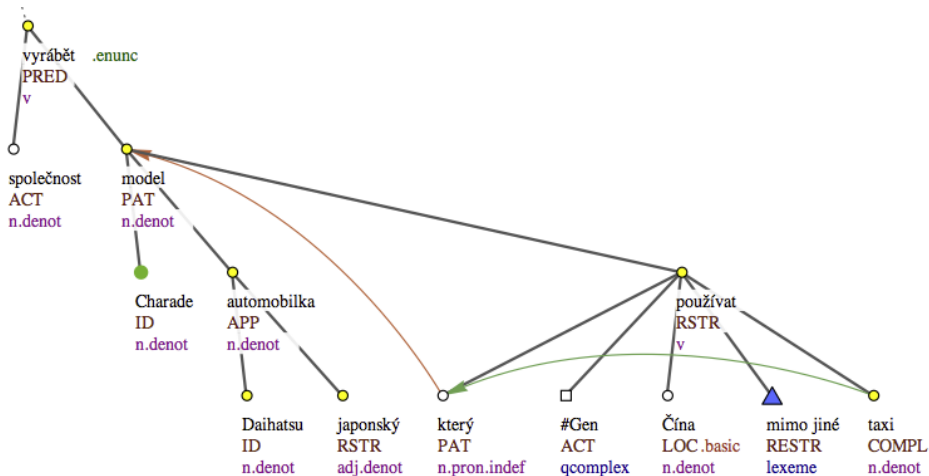
More examples of treebanks

- *Prague Dependency Treebank* – Czech: 1.5 MW
- *Tiger* – German: 0.9 MW
- *LASSY* – Dutch: 1500 MW
- *Lingo Redwoods* – English: 45 KS
- *BulTreeBank* – Bulgarian: 250 KW
- *INESS Treebanking Infrastructure* – various: [Rosén et al.(2012)]
- *Składnica – Polish Constituency Treebank*: 8 KS
- ...

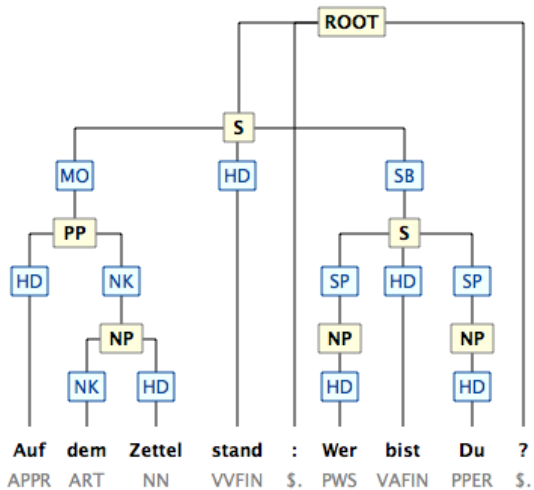
PDT – analytical layer



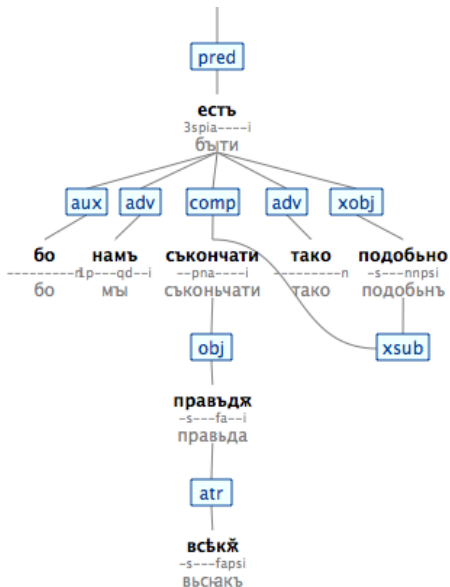
PDT – tectogrammatical layer



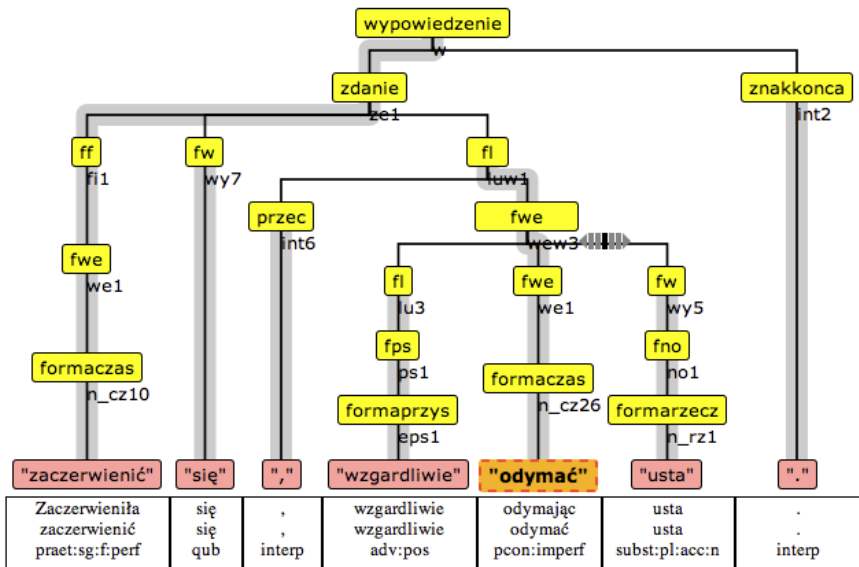
Tiger



Old Church Slavonic (INESS)



Polish (Składnica)



Outline of the talk

- 1 Why treebanks, why grammars?
- 2 Treebanks
- 3 Grammars**
- 4 The grammar–treebank relationship
- 5 Czech treebanking
- 6 Architecture
- 7 Examples
- 8 Input processing
- 9 Conclusions and plans

About grammars

- Treebank grammars [Charniak & Charniak(1996)]
 - Probabilistic grammars directly projected from treebanks
 - “a paradigm shift from the manually constructed, a priori fixed linguistic grammars” [Prescher et al.(2006)]
- Annotation manuals
- Symbolic (rule-based) grammars

The paradigm shift

- Analytical, linguistic
×
empirical, data-driven
- Analytical = analysis of linguistic competence
- Poor coverage → discontinue 'deep' processing?

Anyone need grammars? (Stephan Oepen, TLT2) 1/2

The Ultimate Grammar

- Coverage of arbitrary data, cross-domain and cross-genre
- Adequate grammatical analyses in all cases
- Inclusion of semantics
- Fully declarative
- Same grammar for both parsing and generation
- High-efficiency processing tools

BUT:

- No generally accepted linguistic theory
- Long, tedious, error-prone engineering process
- Few experts

Anyone need grammars? (Stephan Oepen, TLT2) 2/2

The Final Treebank

- Representative data for ‘all’ of the language, domains, and genres
- Full annotation with (at least) syntactic and semantic information
- Utterly coherent
- Free of errors
- Fully documented
- Freely available

BUT:

- No generally accepted annotation standard
- Long, tedious, error-prone annotation process

The answer: grammars and treebanks should go together

- Treebank annotation is where a grammar and a treebank can meet
- Treebank annotation is also where multiple theories can meet and complement each other
- Grammar and treebank are like two sides of a coin:
 - competence × performance
 - system × use
 - langue × parole
 - theoretical × empirical

Outline of the talk

- 1 Why treebanks, why grammars?
- 2 Treebanks
- 3 Grammars
- 4 The grammar–treebank relationship**
- 5 Czech treebanking
- 6 Architecture
- 7 Examples
- 8 Input processing
- 9 Conclusions and plans

Treebank – grammar/theory relations

A treebank is useful ...

- As a source and testbed for grammar/theory development [Hajičová & Sgall(2006)]
- As training data for treebank grammars and other NLP tools

A grammar/theory is useful ...

- To guide the design of an annotation scheme
- To control annotation consistency
- To generate treebank annotations

Linking lexicon and treebank

- Theoretically motivated design
- Start: independently compiled list of entries
- Incremental development

Examples:

- *PDT-VALLEX* [Hajič et al.(2003)]
- *FrameNet* [Palmer et al.(2005)]
- *PropBank* [Baker et al.(1998)]
- *TüBa-D/Z Valency Lexicon* [Hinrichs & Telljohann(2009)]
- ...

Linking grammar and treebank

- Grammar development should be supported by an annotated corpus
- Automatic annotation by symbolic grammars requires a fully adequate grammar, ideally based on a corpus
- Vicious circle? A possible answer: Incremental development of both the grammar and the treebank

Examples:

- *LinGO Redwoods* [Oepen et al.(2002)]
- *Norgram* [Rosén et al.(2006)]
- *BulTreeBank* [Simov et al.(2002)]
- *Składnica* [Świdziński & Woliński(2010)]
- ...

Rarely a single correct parse of a sentence

- Symbolic grammars have limited access to context and world knowledge
- They produce many parses due to morphosyntactic and structural ambiguities

Solutions

- Stochastic disambiguation
- Stochastic ranking
- Manual selection, preferably interactive, based on discriminants

Never 100% coverage

- A parsed corpus generated by a symbolic grammar will never reach 100% coverage of real-world data (LinGO: about 80%)
- Reasons are fundamental: competence \times performance

Some examples:

- anacoluth
- contamination
- attraction
- zeugma
- some cases of extraction

Examples of suboptimal syntax

- (1) Kdo přijde pozdě, nic mu nedají.
 who comes late nothing him not-give
 Who comes late won't get anything. (intended)
- (2) Včera jsem viděl a mluvil s tím člověkem.
 yesterday AUX saw and spoke with that man
 I saw and spoke to that person yesterday.
- (3) Nebo já Gazda nevím, jak diktuje.
 or I Gazda not-know how dictates
 Or I don't know how Gazda dictates. (int'd, due to Jan Klaška)

Beyond grammar

- How to find negative evidence in standard corpora?
- Except for non-words not easy in a corpus of written language
- Much of ‘suboptimal’ language use in spoken and learner corpora
- Grammar useful to detect ungrammaticality
- A treebank of suboptimal German [Kepser et al.(2004)]
- Phenomena-oriented corpus [Oliva(2008)]

Can we build a grammar-based treebank that includes real language?

Possible solutions?

- A combination of stochastic + symbolic methods
- Two grammars: positive and negative [Oliva & Petkevič(1998)]
- Competence + performance grammar [Kempen & Harbusch(2001)]

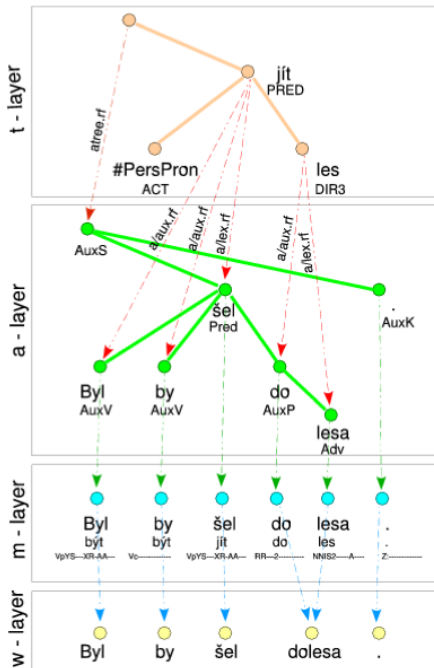
Outline of the talk

- 1 Why treebanks, why grammars?
- 2 Treebanks
- 3 Grammars
- 4 The grammar–treebank relationship
- 5 Czech treebanking**
- 6 Architecture
- 7 Examples
- 8 Input processing
- 9 Conclusions and plans

The treebank of Czech

Prague Dependency Treebank

- Dependency syntax, close to the Prague theory of Functional Generative Description [Sgall et al.(1986)]
- 3 annotation levels: morphology, surface syntax, deep syntax
- PDT 0.5 – 1998, 0.5 MW
- PDT 1 – 2000, 1.5 MW
- PDT 2 – 2004, deep syntax
- PDT 2.5 – 2011, multi-word units, clause segmentation



Time to scale up?

- 1.5 MW still too few for investigating less frequent forms and phenomena
 - Could offer more annotation formats
 - Could support inherent syntactic ambiguities
- (4) Přinesl bednu ze sklepa.
brought box from cellar
He brought a box from the cellar
- (5) krajíc chleba s máslem
slice bread with butter
a buttered slice of bread

A treebank for every taste

Theory-Supporting Treebank [Nivre(2003)]

- **Theory-neutral** annotation contains too little information or too many compromises to be really useful
- **Theory-specific** may shut out people from other research traditions
- **Conversion?** But the source annotation often lacks information to support a completely accurate conversion.
- Possible conversions as a requirement in the design of treebank annotation schemes. Different kinds of (theory-specific) annotation should be supported by an **underlying internal representation**.

A treebank for every taste

Multi-Representational Treebank [Xia et al.(2009)]

- **Definitional differences** between phrase structure and dependency structure: convertible if designed properly.
- **Preferential differences** – the same in both: empty categories; labels to edges; ordered or unordered trees.

Can a single core annotation be viewed in different ways?

- Theory-specific representations have different appearances but share a large part of content: constituency/dependency, morphosyntactic categories, even the spirit of analyses of many phenomena
- A treebank offering different views of a sufficiently expressive annotation scheme is a realistic goal
- Additional benefit: relating linguistic theories

A larger treebank with customizable visualization?

Short-term goals:

- Syntactic annotation of the Czech National Corpus (1.3 billion words) using a stochastic parser, followed by a rule-based correction module
- Robust and expressive core annotation format, potentially underspecified
- Customizable query, visualization and export interface, offering multiple options to view syntactic structure
- Accessible to lay users and satisfying experts at the same time

Long-term goals:

- Development of a corpus-based grammar
- Options for queries, visualization and export:
 - ready-made, tailored to specific theories, or
 - defined by the user
- Development of the correction module

The tasks of the grammar

- Checking consistency
- Adding more information on top of existing annotation
- Assisting the treebank user
- To help converting the data onto other formats more easily
- To help distinguishing grammatical and suboptimal/ungrammatical forms and structures

Grammar design and development

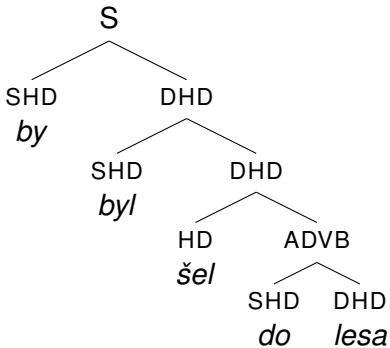
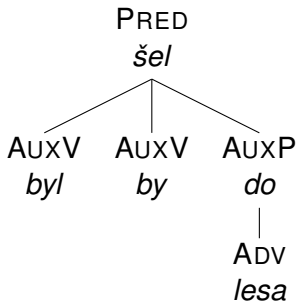
- Constraint-based: all is possible except when stipulated otherwise
- Hand-crafted but verified against the corpus data
- Incremental development, based on conversion rules
- Underspecification, partial parses to cope with suboptimal/ungrammatical forms and constructions
- Performance grammar as a mediator with the real-world language, similar to negative grammar?

Outline of the talk

- 1 Why treebanks, why grammars?
- 2 Treebanks
- 3 Grammars
- 4 The grammar–treebank relationship
- 5 Czech treebanking
- 6 Architecture**
- 7 Examples
- 8 Input processing
- 9 Conclusions and plans

Syntactic structure

- Internal skeleton structures: constituency-based, with a combination of **binary** and **flat** branching
- Interpretable as **constituency** or **dependency** trees, according to users' specification, visualized with an arbitrary amount of detail, not necessarily by tree graphs
- Surface and deep structure encoded within a single structure: constituents are labelled as **syntactic functions** including **head** as a special function
- Heads are further specified as **deep** or **surface**
 - **Deep head**: deep syntactic governor: *bylo by se to povedlo*
 - **Surface head**: can be identical to the deep head or different: auxiliary, prepositions, subordinate conjunctions, numerals



Three levels

- Word order and syntactic structure as distinct dimensions, each sentence is represented at three inter-linked levels:
 - **graphemics** (orthographic words, contractions)
 - **morphology** (syntactic words, including haplogitized items)
 - **syntax** (trees, no nodes for pro-dropped subjects)

Annotation of syntactic phenomena

- Agreement of various types
- Compound periphrastic verbal forms (passives, conditional structures, future...)
- Grammatical co-reference (grammatical control, relative/reflexive pronouns, predicative complements)
- Multi-word units (collocations)

Expressive power

- Expressive enough to accommodate analyses of arbitrary granularity
- Ambiguous or undecidable phenomena represented by underspecification and distributive disjunction
- Annotation of any kind can be missing, a sentence may be a mere list of words

Specifications

- Annotation must be licensed by a formal grammar. Words and constituents have their appropriate (potentially underspecified) sets of features
- Lexicons are used to index forms, syntactic words and compound forms
- Customizable visualizations are enabled by formal definitions

Links within a tree

- Agreement
- Compound (multi-word) verbal predicates
- Grammatical coreference
- ...

Syntactic structure

- each nonterminal node is assigned a construction type and a syntactic function
- each terminal node is assigned a syntactic function

Hierarchy of construction types

- **Headed**
- **UnHeaded**
 - **Coord** – coordination
 - **Adord** – adordination
 - **Unspec** – unspecified (for collocations and other)

Function for **UnHeaded** structures:

- **Memb** – a member

Syntactic functions for **Headed**

- **SurfHead** – surface head: auxiliary *být/bývat*, prepositions, subordinate conjunctions, numerals in quantified expressions: *pět dětí*
- **DeepHead** – in case it differs from SurfHead (head nouns in PPs, autosemantic verbs in analytical predicates...)
- **Head** – both **SurfHead** and **DeepHead**

Other syntactic functions for **Headed**

- **Subj** – subject
- **Attr** – attribute
- **Obj-Advb**
 - **Obj**
 - **Advb**
- **VbAttr** – predicative complement
- **RefITant** – reflexive element (*si, se*) for inherent reflexives
- **Deagent** – deagentive reflexive
- **Apos** – apposition
- **InDep** – independent syntactic element (parenthesis, vocative syntactic noun...)

Outline of the talk

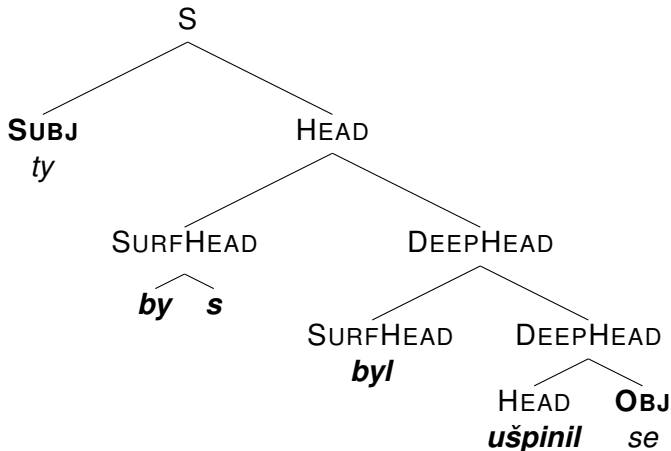
- 1 Why treebanks, why grammars?
- 2 Treebanks
- 3 Grammars
- 4 The grammar–treebank relationship
- 5 Czech treebanking
- 6 Architecture
- 7 Examples**
- 8 Input processing
- 9 Conclusions and plans

Treating contractions

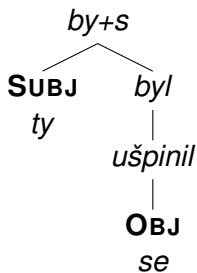
- (6) **Ty** by **ses** byl ušpinil.
 you would REFL+AUX_{2nd,sg} be_{pple} get dirty_{pple}
 ‘You would have got dirty.’

Ty by se byl ušpinil.

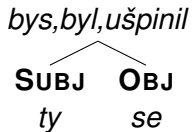
(7)



- (8) **Surface dependency structure** derived from (7)



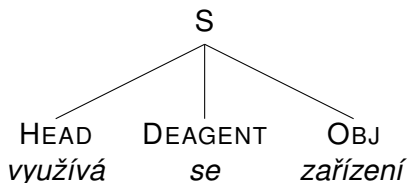
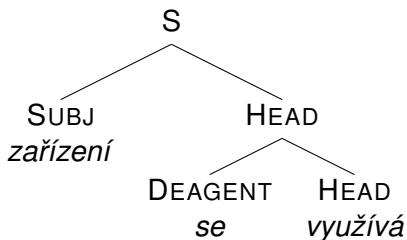
- (9) **Deep dependency structure** derived from (7)



Subject/object ambiguity

Reflexive passive:

- (10) Zařízení_{Nom/Gen} se využívá.
 device REFL uses
 'The device is being used.'

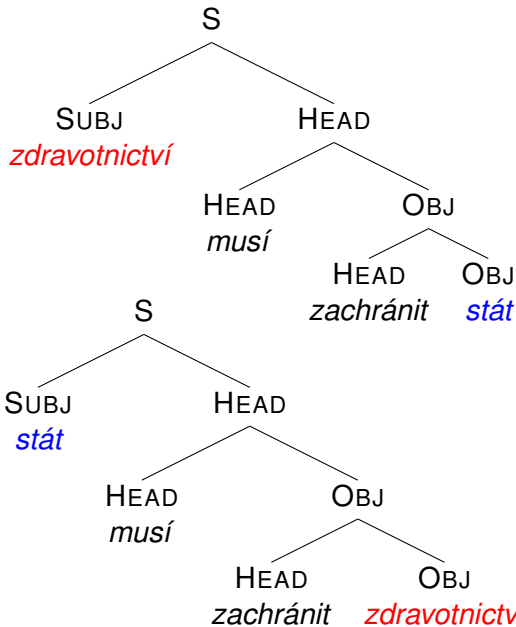


Another type of subject/object ambiguity

- (11) **Zdravotnictví** musí zachránit **stát**.
health service_{nom/acc} must save **state**_{nom/acc}

Two different readings:

- #1 Health service must save the State.
- #2 Health service must be saved by the government.



Outline of the talk

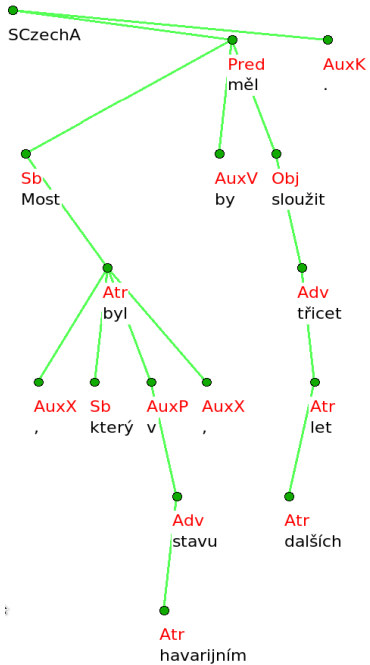
- 1 Why treebanks, why grammars?
- 2 Treebanks
- 3 Grammars
- 4 The grammar–treebank relationship
- 5 Czech treebanking
- 6 Architecture
- 7 Examples
- 8 Input processing**
- 9 Conclusions and plans

Processing of the input text:

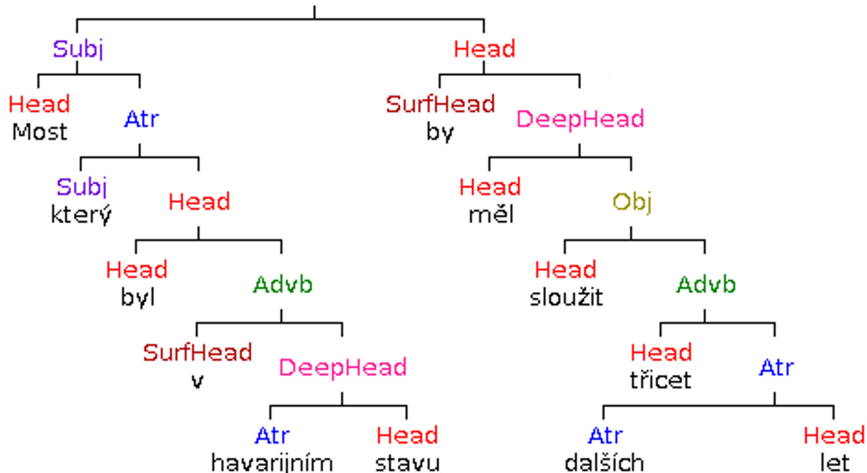
- Automatic correction of the output of a stochastic parser
- Conversion of the corrected parse + modifications:
 - phenomena that require arbitrary decisions in a dependency tree: constructions with function words, coordinated constructions, lists
 - disjunction accounting for structural ambiguities expressed by PDT's "combined functions" AttrAdv, ObjAdv

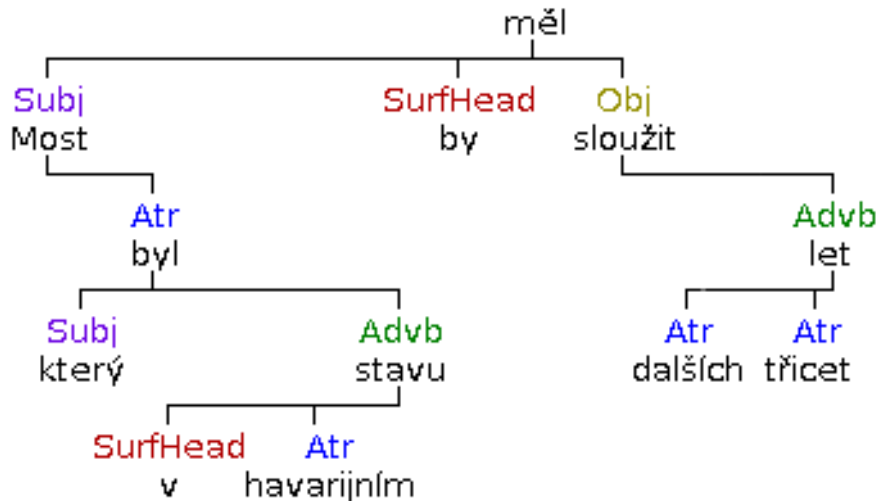
Syntactic tree in the PDT and the new format

- (12) Most, který byl v havarijním stavu, by měl sloužit
 Bridge which was in emergency state should have_{modal} serve
 dalších třicet let.
 next thirty years.
 ‘The bridge, which was ramshackle, should serve for another
 thirty years.’



Most , který byl v havarijním stavu , by měl sloužit dalších třicet let .





Correction module

- 30 correction rules so far
- For more frequent errors which can be reliably corrected
- Such as noun in accusative as subject

Success rate of the correction modules

	Rules	Dependency	Label	Total
Clauses	6	1688	774	1744
NP	8	819	2066	2625
PP	9	834	7160	7722
Other	5	412	1390	1802
Total (ppm)		3753	11390	13893
Total (%)		0.38%	1.14%	1.39%

Outline of the talk

- 1 Why treebanks, why grammars?
- 2 Treebanks
- 3 Grammars
- 4 The grammar–treebank relationship
- 5 Czech treebanking
- 6 Architecture
- 7 Examples
- 8 Input processing
- 9 Conclusions and plans**

Conclusions and plans 1/2

Results

- Conversion rules
- Correction module
- 200M corpus parsed and corrected
- Beta version of a viewer with three representation modes

Further work

- Manually tagged and parsed subcorpus will provide better data to train the parser
- More parsing errors will be detected and corrected
- More modes of viewing the syntactic structure
- Grammar development

Conclusions and plans 2/2

Empiricism and theory meet in the corpus annotation

- Competence grammar to fully license the annotation of grammatical forms and constructions
- Underspecification and partial parses for the rest
- Performance grammar to close the gap between the real language and the annotation provided by the competence grammar

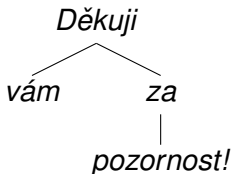
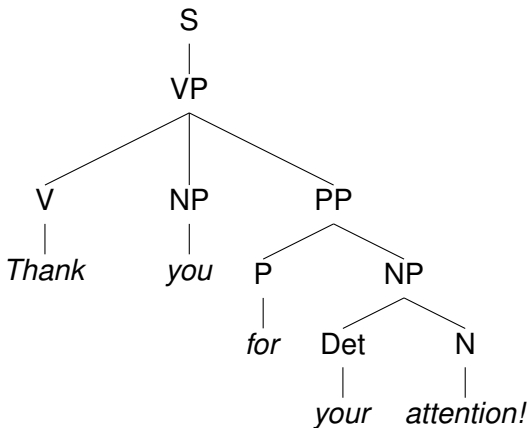
Based on the work of:

Milena Hnátková, Petr Jäger,
Tomáš Jelínek, Vladimír Petkevič,
Hana Skoumalová and myself

Supported by:

The Grant Agency of the Czech Republic

Project no. GAČR P406/10/0434



References I



Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998).

The Berkeley FrameNet project.

In 36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98), pages 86–90, Montréal.



Charniak, E. & Charniak, E. (1996).

Tree-bank grammars.

In In Proceedings of the Thirteenth National Conference on Artificial Intelligence, pages 1031–1036.



Haider, H. (1993).

Deutsche Syntax – Generativ.

Narr, Tübingen.

References II



Hajič, J., Panevová, J., Urešová, Z., Bémová, A., & Pajas, P. (2003).

PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation.

In [Proceedings of The Second Workshop on Treebanks and Linguistic Theories](#), pages 57–68. Växjö University Press.



Hajičová, E. & Sgall, P. (2006).

Corpus annotation as a test of a linguistic theory.

In [Proceedings of LREC 2006](#), pages 879–884.



Hinrichs, E. W. & Telljohann, H. (2009).

Constructing a valence lexicon for a treebank of German.

In [Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories](#), page 41–52.

References III



Kempen, G. & Harbusch, K. (2001).

Performance grammar: a declarative definition.

In M. Theune, A. Nijholt, and H. Hondorp, editors, CLIN, volume 45 of Language and Computers – Studies in Practical Linguistics, pages 148–162. Rodopi.



Kepser, S., Steiner, I., & Sternefeld, W. (2004).

Annotating and querying a treebank of suboptimal structures.

In In Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories (TLT2004), pages 63–74.





Nivre, J. (2003).

Theory-supporting treebanks.

In Proceedings of the Second Workshop on Treebanks and Linguistic Theories.

References IV

-  Oepen, S., Flickinger, D., Toutanova, K., & Manning, C. D. (2002). LinGO Redwoods: A rich and dynamic treebank for HPSG. In Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02), Sozopol, Bulgaria.
-  Oliva, K. (2008). Phenomena-oriented corpora: a manifesto. In F. Štícha and M. Fried, editors, Grammar & Corpora = Gramatika a korpus 2007. Sborník příspěvků ze stejnojmenné konference 25.-27. 9. 2007, Liblice= Selected contributions from the conference Grammar and Corpora, Sept. 25-27, 2007, Liblice, pages 77–104, Praha. Academia.

References V



Oliva, K. & Petkevič, V. (1998).

Phenomena-based description of dependency-syntax: A survey of ideas and formalization.

In E. Hajičová and B. Hladká, editors, Issues of Valency and Meaning – Studies in Honour of Jarmila Panevová. Charles University Press, Praha.



Palmer, M., Gildea, D., & Kingsbury, P. (2005).

The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, **31**(1), 71–106.



Prescher, D., Scha, R., Sima'an, K., & Zollmann, A. (2006).

What are treebank grammars?

In BNAIC'06: BeNeLux conference on Artificial Intelligence 2006, Namur, Belgium.

References VI



Rosén, V., de Smedt, K., & Meurer, P. (2006).

Towards a toolkit linking treebanking to grammar development.

In Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories (TLT'05), Prague, Czech Republic.



Rosén, V., Smedt, K. D., Meurer, P., & Dyvik, H. (2012).

An open infrastructure for advanced treebanking.

In J. Hajič, K. D. Smedt, M. Tadić, and A. Branco, editors, Proceedings of the META-RESEARCH Workshop on Advanced Treebanking, LREC 2012, pages 22–29, Istanbul, Turkey. ELRA, European Language Resources Association.

References VII



Sgall, P., Hajičová, E., & Panevová, J. (1986).
The Meaning of the Sentence in its Semantic and Pragmatic Aspects.

Reidel and Academia, Dordrecht and Praha.
Editor: Jacob Mey.



Simov, K., Osenova, P., Kolkovska, S., Balabanova, E., Doikoff, D., Ivanova, K., & Alexander Simov, M. K. (2002).

Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank.

In Proceedings of LREC 2002, pages 1729–1736, Canary Islands, Spain.

References VIII



Xia, F., Rambow, O., Bhatt, R., Palmer, M., & Sharma, D. M. (2009).

Towards a multi-representational treebank.

In F. Van Eynde, A. Frank, G. van Noord, and K. De Smedt, editors, Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7), pages 127–133, Utrecht. LOT.



Świdziński, M. & Woliński, M. (2010).

Towards a bank of constituent parse trees for Polish.

In Proceedings of the 13th International Conference on Text, Speech and Dialogue, TSD'10, pages 197–204, Berlin, Heidelberg. Springer-Verlag.