

Language Resources and Machine Learning

Sašo Džeroski

Department of Knowledge Technologies

Institut Jožef Stefan, Ljubljana, Slovenia

<http://www-ai.ijs.si/SasoDzeroski/>

Talk outline

- Language technologies and linguistics
- Language resources
- The Multext-East resources
 - Learning morphological analysis/synthesis
 - Learning PoS tagging
 - Lemmatization
- The Prague Dependency Treebank
 - Learning to assign tectogrammatical functors

Language Technologies – Apps.

- Machine translation
- Information retrieval and extraction, text summarisation, term extraction, text mining
- Question answering, dialogue systems
- Multimodal and multimedia systems
- Computer assisted: authoring; language learning; translating; lexicology; language research
- Speech technologies

Linguistics: The background of LT

What is language?

- *Act of speaking* in a given situation
- The individual's system underlying this act
- The *abstract system* underlying the collective totality of the speech/writing behaviour of a community
- The *knowledge of this system* by an individual

What is linguistics?

- The scientific study of language
- General, theoretical, formal, mathematical, computational linguistics

Comp Ling = The computational study of language

- Cognitive simulation; Natural language processing

Levels of linguistic analysis

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Discourse analysis
- Pragmatics

- + Lexicology

Morphology

- The study of the structure and form of words
- Morphology as the interface between phonology and syntax (and the lexicon)
- Inflectional and derivational (word-formation) morphology
- Inflection (syntax-driven):
gledati, gledam, gleda, glej, gledal,...
- Derivation (word-formation):
pogledati, zagledati, pogled, ogledalo,...,
zvezdogled (compounding)

Inflectional morphology

- Mapping of form to (syntactic) function
- *dogs* -> *dog* + *s* / DOG [N,pl]
- In search of regularities: *talk/walk*;
talks/walks; *talked/walked*; *talking/walking*
- Exceptions: *take/took*, *wolf/wolves*,
sheep/sheep
- English (relatively) simple; inflection much richer in, e.g., Slavic languages

Syntax

- How are words arranged to form sentences?
- **I milk like*
- *I saw the man on the green hill with a telescope.*
- The study of rules which reveal the structure of sentences (typically tree-based)
- A “pre-processing step” for semantic analysis
- Terms: Subject, Object, Noun phrase, Prepositional phrase, Head, Complement, Adjunct,...

Semantics

- The study of *meaning* in language
- Very old discipline, esp. philosophical semantics (Plato, Aristotle)
- Under which conditions are statements true or false; problems of quantification
- Terms: Actor, Conjunction, Patient, Predicate

- The meaning of words – lexical semantics
- *spinster* = unmarried female
 - **My brother is a spinster*

Lexicology

- The study of the vocabulary (lexis / lexemes) of a language (a lexical “entry” can describe less or more than one word)
- Lexica can contain a variety of information: sound, pronunciation, spelling, syntactic behaviour, definition, examples, translations, related words
- Dictionaries, digital lexica
- Play an increasingly important role in theories and computer applications
- Ontologies: WordNet, Semantic Web

Computational Linguistics

Processes, methods and resources

- The Oxford Handbook of Computational Linguistics
 - Edited by R. Mitkov, ed.
- Processes: **Text-to-Speech Synthesis; Speech Recognition; Text Segmentation; Part-of-Speech Tagging; Lemmatisation; Parsing; Word-Sense Disambiguation; Anaphora Resolution; Natural Language Generation**
- Methods: **Finite-State Technology; Statistical Methods; Machine Learning; Lexical Knowledge Acquisition**
- Resources: **Lexica; Corpora; Ontologies**

Language Resources/Corpora

- Lexica (lexicon), corpora (corpus), ontologies (e.g. WordNet)
- A corpus is a collection or body of writings/texts
- EAGLES (Expert Advisory Group on Language Engineering Standards) definition: a corpus is
 - a collection of pieces of language
 - that are selected and ordered according to explicit linguistic criteria in order
 - to be used as a sample of the language
- A computer corpus is encoded in a standardised and homogeneous way for open-ended retrieval tasks

The use of corpora

Corpora can be annotated at various levels of linguistic analysis (morphology, syntax, semantics)

Lemmas (M), parse trees/dependency trees (Syn), TG trees (Sem)

Corpora can be used for a variety of purposes. These include

- Language learning
- Language research (descriptive linguistics, computational approaches, empirical linguistics)
 - lexicography (mono/bi-lingual dictionaries, terminological)
 - general linguistics and language studies
 - translation studies

We can use corpora for the development of LT methods

- as testing sets for (manually) developed methods
- as training sets to (automatically) develop methods with ML

Corpora Annotation: Morphology

```
<s id="0sl.1.2.3.4">
  <w lemma="Winston" ana="Npmsn">Winston</w>
  <w lemma="se" ana="Px-----y">se</w>
  <w lemma="biti" ana="Vcip3s--n">je</w>
  <w lemma="napotiti" ana="Vmpps-sma">napotil</w>
  <w lemma="proti" ana="Spsd">proti</w>
  <w lemma="stopnica" ana="Ncfpd">stopnicam</w>
  <c>.</c>
</s>
```

Winston made for the stairs.

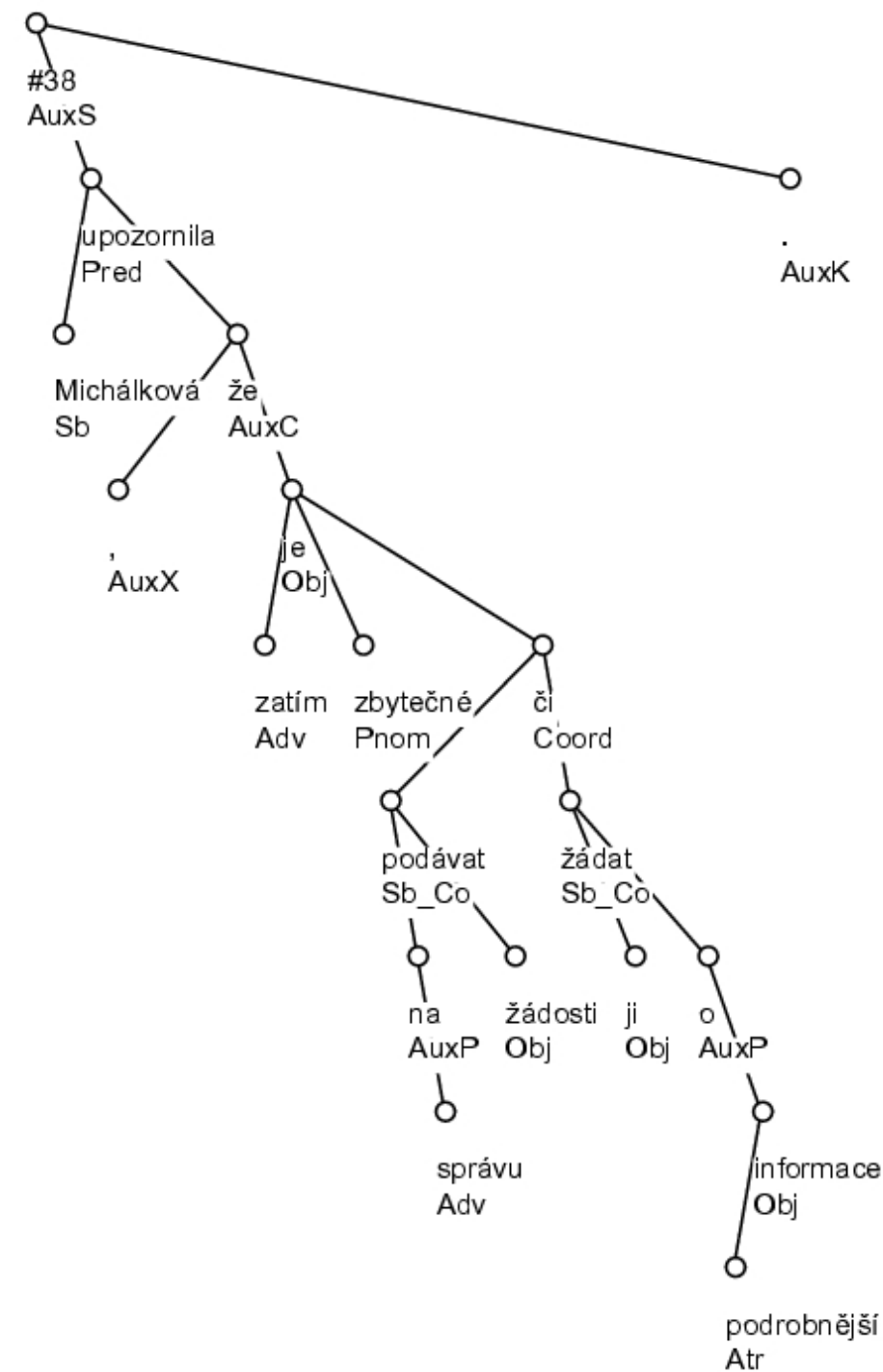
Winston se je napotil proti stopnicam.

CORPORA ANNOTATION: SYNTAX

Michalkova upozornila, že zatím je
zbytečně podávat na správu žádosti
či žádat ji o podrobnější informace.

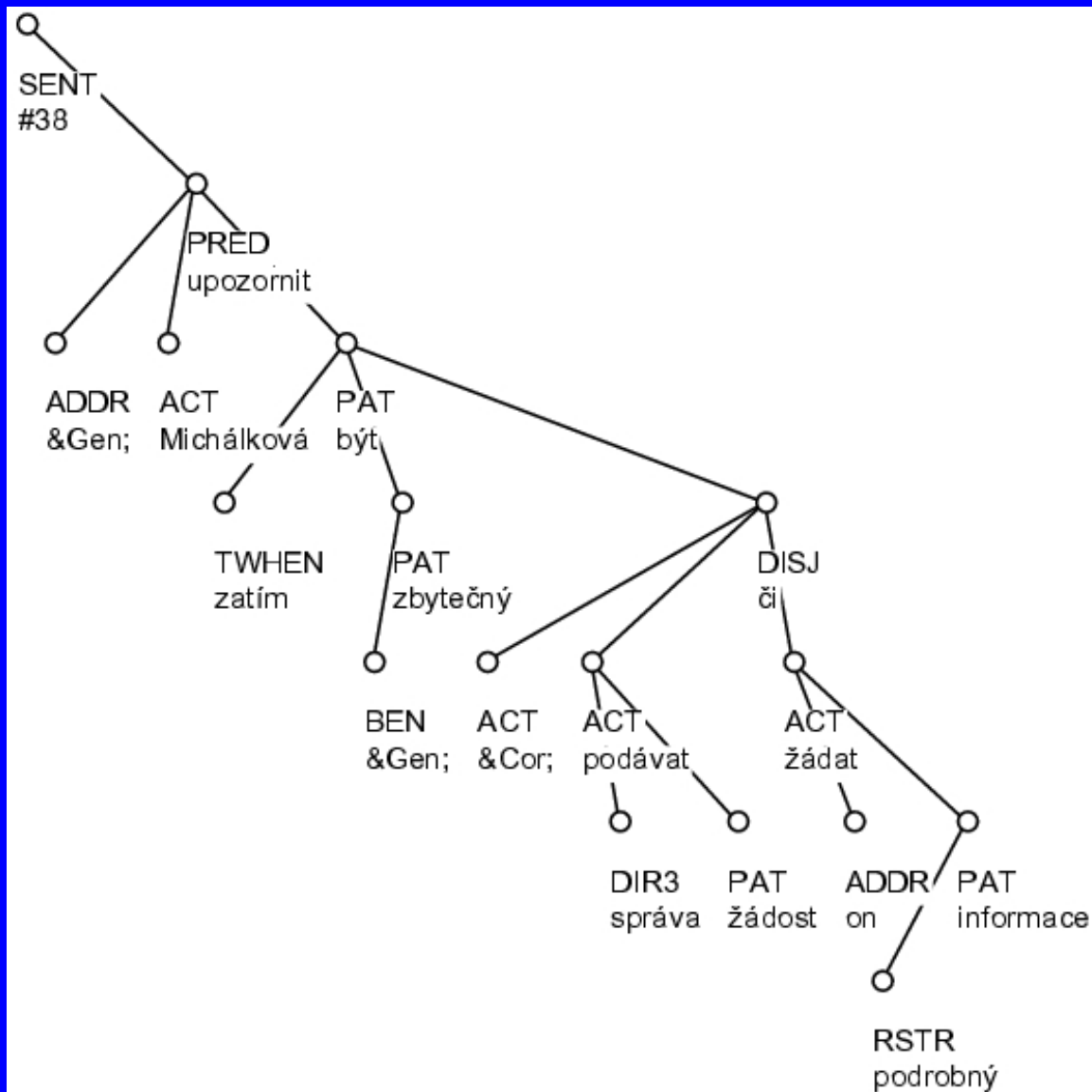
Literal translation:

Michalkova pointed-out that meanwhile
is superfluous to-submit to administration
requests or to-ask it
for more-detailed information.



CORPORA ANNOTATION: SEMANTICS

“M. pointed out that for the time being it was superfluous to submit requests to the administration, or to ask it for more detailed information.”



Literal translation:

Michalkova pointed-out
that meanwhile

is superfluous to-submit
to administration requests

or to-ask it

for more-detailed information.

Talk outline

- Language technologies and linguistics
- Language resources
- The Multext-East resources
 - Learning morphological analysis/synthesis
 - Learning PoS tagging
 - Lemmatization
- The Prague Dependency Treebank
 - Learning to assign tectogrammatical functors

MULTEXT-East COPERNICUS Project

- Multilingual Text Tools and Corpora for Central and Eastern European Languages
- Produced corpora and lexica for
 - Bulgarian (Slavic)
 - Czech (Slavic)
 - Estonian (Finno-Ungric)
 - Hungarian (Finno-Ungric)
 - Romanian (Romance)
 - **Slovene** (Slavic)
- Results published on CD-ROM
- CD-ROM mirror and other information on the project can be found at <http://nl.ijs.si/ME/>

MULTEXT-East Home Page



 **Multext-East Home Page**

MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages

The MULTEXT-East resources are a multilingual dataset for language engineering research and development. This dataset contains, for Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Resian, Romanian, Russian, Slovene, and Serbian, some, or all of the following language resources: the MULTEXT-East morphosyntactic specifications, lexica, and annotated "1984" corpus; the MULTEXT-East parallel and comparable text and speech corpora; and associated documentation.

New: [MULTEXT-East resources Version 3](#) (latest release: 2004-07-07)

What's new in V3:

- all corpora now encoded in XML [TEI P4](#)
- joins together the resources from [Version 1](#) (1998) and [Version 2](#) (2002)
- adds [Serbian annotated "1984"](#) and [Resian morphosyntactic specifications](#)
- an updated [bibliography](#)
- many errors from previous versions corrected...

MULTEXT-East 1984 corpus

2.3. MULTEXT-East 1984 corpus

The novel "1984" by George Orwell is the central component of the MULTEXT-East corpus. This parallel corpus annotated contains the novel in the English original (about 100,000 words in length), and its translations into a number of languages.

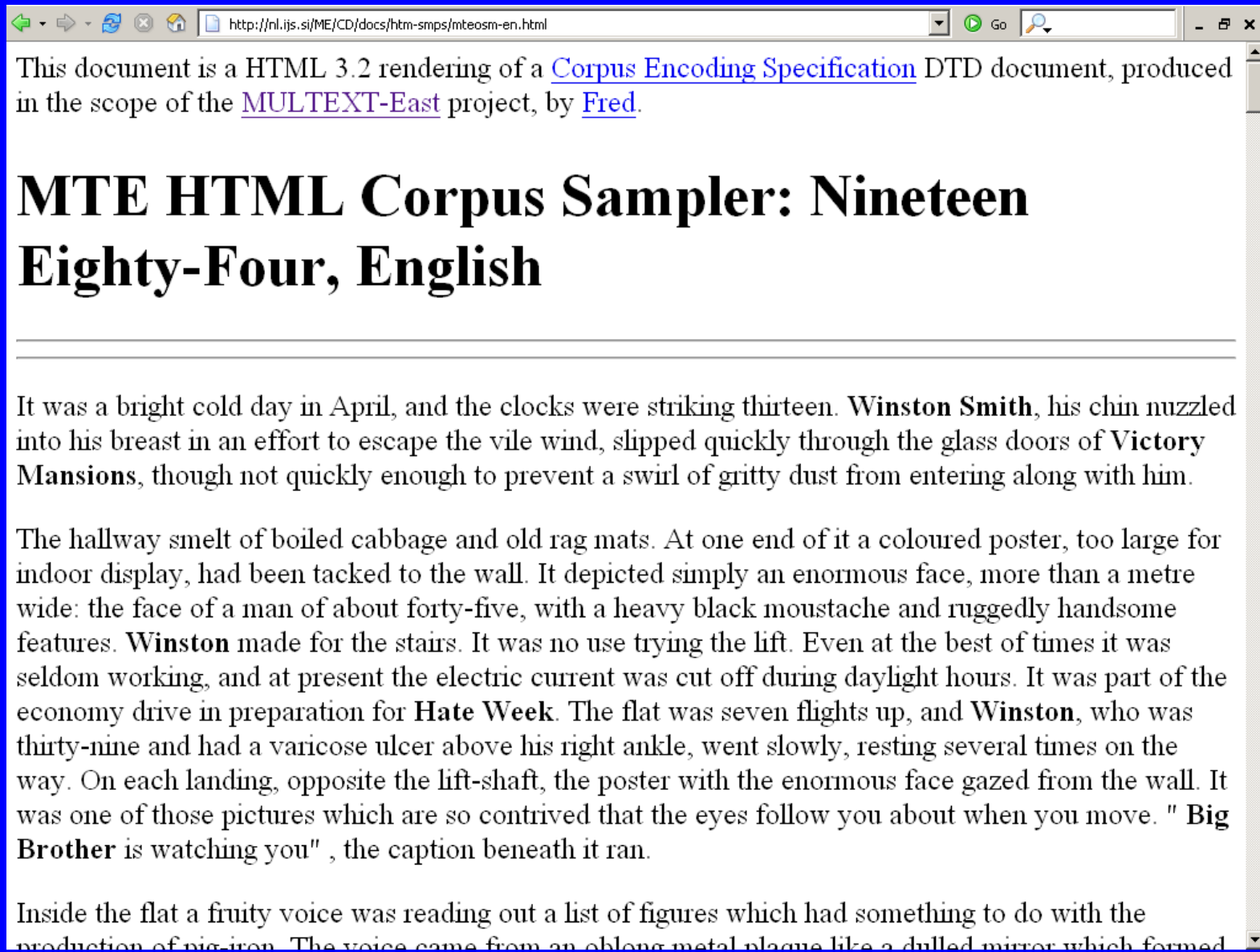
George Orwell
Nineteen
Eighty-Four



It was an enormous pyramidal structure of glittering white concrete, soaring up, terrace after terrace, 300 metres into the air. From where **Winston** stood it was just possible to read, picked out on its white face in elegant lettering, the [three slogans](#) of the Party:

War is peace	Freedom is slavery	Ignorance is strength
Războiul este pace	Libertatea este sclavie	Ignoranța este putere
Vojna je mir	Svoboda je suženjstvo	Nevednost je moč
Válka je mír	Svoboda je otroctví	Nevědomost je síla
Воїната е мир	Свободата е робство	Невежеството е сила
Sõda on rahu	Vabadus on orjus	Teadmatus on jõud
Rat je mir	Sloboda je ropstvo	Neznanje je moć
A háború: béke	A szabadság: szolgaság	A tudatlanság: erő
Karas — tai taika	Laisvė — tai vergija	Nežinomas — tai jėga
Rat je mir	Sloboda je ropstvo	Neznanje je moć
Воїна — это мир	Свобода — это рабство	Незнание — сила

Corpus Example: Document



This document is a HTML 3.2 rendering of a [Corpus Encoding Specification](#) DTD document, produced in the scope of the [MULTEXT-East](#) project, by [Fred](#).

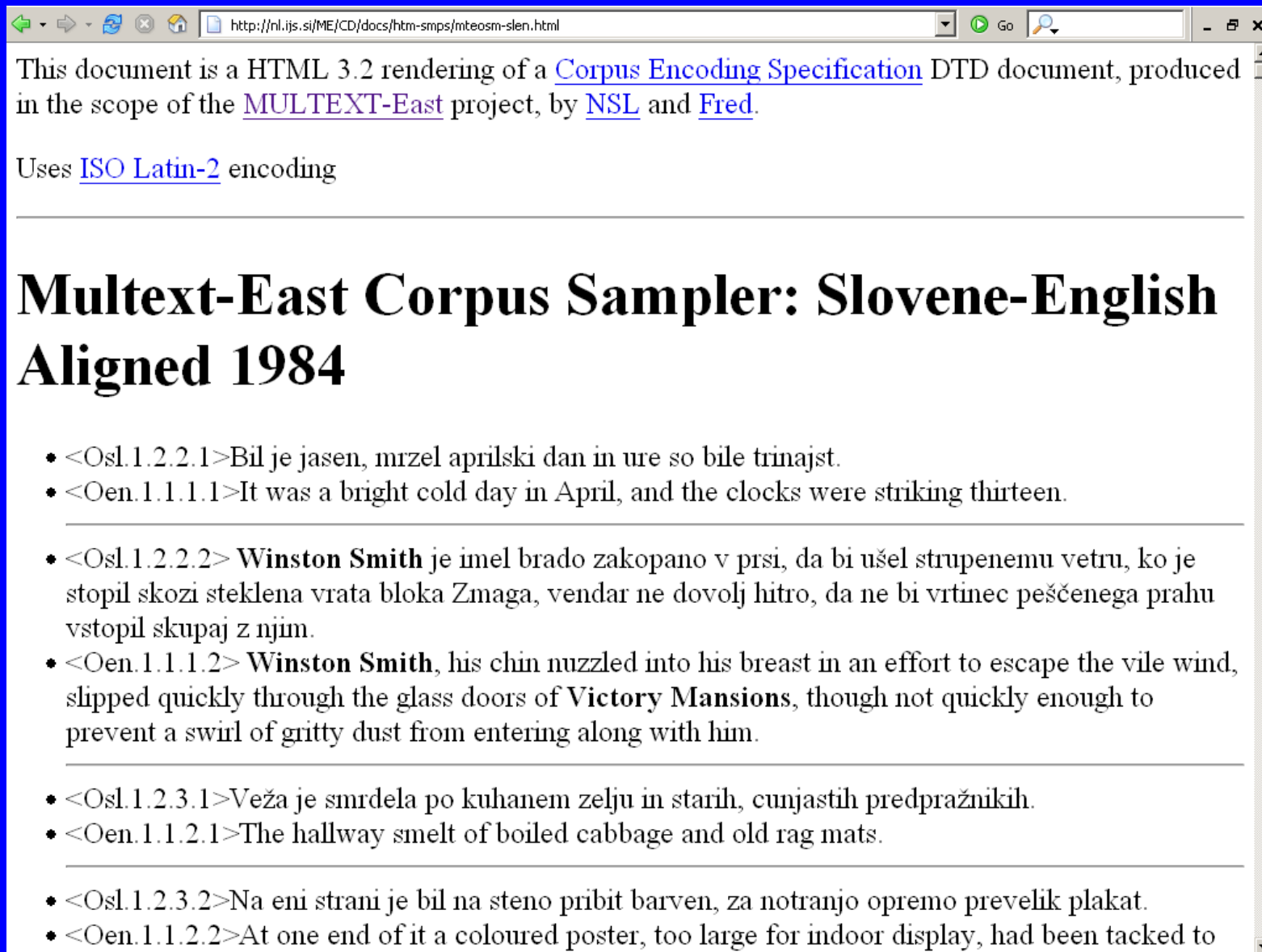
MTE HTML Corpus Sampler: Nineteen Eighty-Four, English

It was a bright cold day in April, and the clocks were striking thirteen. **Winston Smith**, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of **Victory Mansions**, though not quickly enough to prevent a swirl of gritty dust from entering along with him.

The hallway smelt of boiled cabbage and old rag mats. At one end of it a coloured poster, too large for indoor display, had been tacked to the wall. It depicted simply an enormous face, more than a metre wide: the face of a man of about forty-five, with a heavy black moustache and ruggedly handsome features. **Winston** made for the stairs. It was no use trying the lift. Even at the best of times it was seldom working, and at present the electric current was cut off during daylight hours. It was part of the economy drive in preparation for **Hate Week**. The flat was seven flights up, and **Winston**, who was thirty-nine and had a varicose ulcer above his right ankle, went slowly, resting several times on the way. On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall. It was one of those pictures which are so contrived that the eyes follow you about when you move. " **Big Brother** is watching you" , the caption beneath it ran.

Inside the flat a fruity voice was reading out a list of figures which had something to do with the production of pig-iron. The voice came from an oblong metal plaque like a dulled mirror which formed

Corpus Example: Alignment



This document is a HTML 3.2 rendering of a [Corpus Encoding Specification](#) DTD document, produced in the scope of the [MULTEXT-East](#) project, by [NSL](#) and [Fred](#).

Uses [ISO Latin-2](#) encoding

Multext-East Corpus Sampler: Slovene-English Aligned 1984

- <Osl.1.2.2.1> Bil je jasen, mrzel aprilski dan in ure so bile trinajst.
- <Oen.1.1.1.1> It was a bright cold day in April, and the clocks were striking thirteen.

- <Osl.1.2.2.2> **Winston Smith** je imel brado zakopano v prsi, da bi ušel strupenemu vetru, ko je stopil skozi steklena vrata bloka Zmaga, vendar ne dovolj hitro, da ne bi vrtinec peščenega prahu vstopil skupaj z njim.
- <Oen.1.1.1.2> **Winston Smith**, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of **Victory Mansions**, though not quickly enough to prevent a swirl of gritty dust from entering along with him.

- <Osl.1.2.3.1> Veža je smrdela po kuhanem zelju in starih, cunjastih predpražnikih.
- <Oen.1.1.2.1> The hallway smelt of boiled cabbage and old rag mats.

- <Osl.1.2.3.2> Na eni strani je bil na steno prabit barven, za notranjo opremo prevelik plakat.
- <Oen.1.1.2.2> At one end of it a coloured poster, too large for indoor display, had been tacked to

Corpus/Lexicon Example: Tagging

```
<s id="0sl.1.2.3.4">
  <w lemma="Winston" ana="Npmsn">Winston</w>
  <w lemma="se" ana="Px-----y">se</w>
  <w lemma="biti" ana="Vcip3s--n">je</w>
  <w lemma="napotiti" ana="Vmpps-sma">napotil</w>
  <w lemma="proti" ana="Spsd">proti</w>
  <w lemma="stopnica" ana="Ncfpd">stopnicam</w>
  <c>.</c>
</s>
```

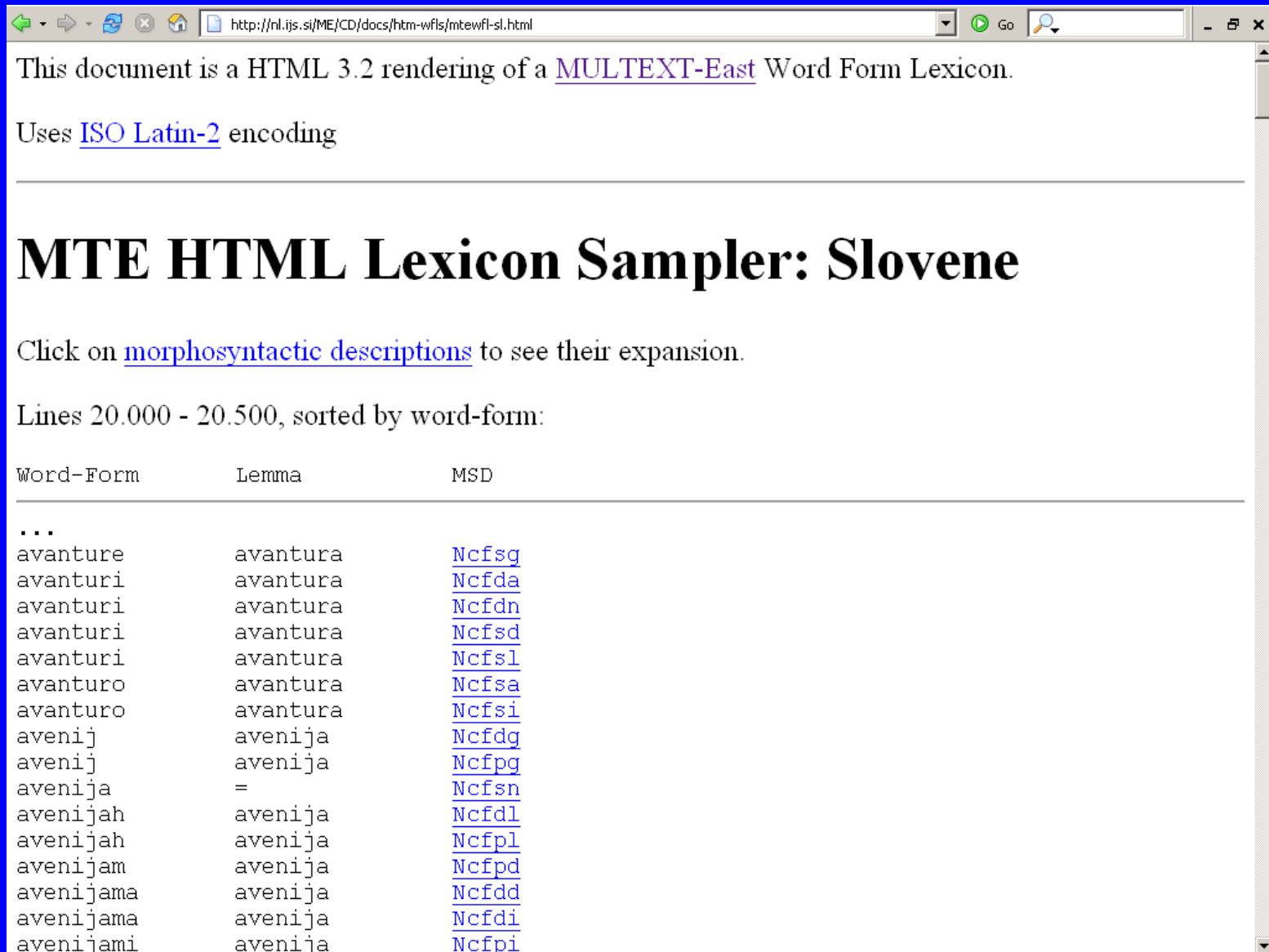
Winston made for the stairs.

Winston se je napotil proti stopnicam.

Slovene Lexicon

- Tabular format
- Covers all inflectional forms of corpus lemmas
- Comprises 560000 entries, 200000 word-forms, 15000 lemmas,
- 2000 MSDs (Morpho-Syntactic Descriptions)
- Morpho-syntactic specifications
 - Categories
 - Noun
 - Verb
 - ...
 - Particle
 - Tables of attribute values

Lexicon Example: Entries



This document is a HTML 3.2 rendering of a [MULTEXT-East](#) Word Form Lexicon.

Uses [ISO Latin-2](#) encoding

MTE HTML Lexicon Sampler: Slovene

Click on [morphosyntactic descriptions](#) to see their expansion.

Lines 20.000 - 20.500, sorted by word-form:

Word-Form	Lemma	MSD
...		
avanture	avantura	Ncfsg
avanturi	avantura	Ncfda
avanturi	avantura	Ncfdn
avanturi	avantura	Ncfds
avanturi	avantura	Ncfsl
avanturo	avantura	Ncfsa
avanturo	avantura	Ncfsi
avenij	avenija	Ncfdg
avenij	avenija	Ncfpg
avenija	=	Ncfsn
avenijah	avenija	Ncfdl
avenijah	avenija	Ncfpl
avenijam	avenija	Ncfpd
avenijama	avenija	Ncfdd
avenijama	avenija	Ncfdi
avenijami	avenija	Ncfpi

Lexicon Example: Grammar

- Noun

```

11 Positions
****  ****  ****  ****  ****  -----  -----  -----  -----  -----
PoS   Type  Gend  Numb  Case  Def    Cltc  Anim  OwnN  OwnP  OwdN
****  ****  ****  ****  ****  -----  -----  -----  -----  -----
=  =====  =====  =  RO  SL  CS  BG  ET  HU
P  ATT          VAL      C  x  x  x  x  x  x
=  =====  =====  =
1  Type          common    c  x  x  x  x  x  x
   proper      p  x  x  x  x  x  x
-  -----  -----  -
2  Gender          masculine  m  x  x  x  x
   feminine      f  x  x  x  x
   neuter        n  x  x  x  x
-  -----  -----  -
3  Number          singular   s  x  x  x  x  x  x
   plural        p  x  x  x  x  x  x
   dual          d    x  x
           l.s.  count      t    x
-  -----  -----  -

```

= =====	=====	=	RO	SL	CS	BG	ET	HU
4 Case	nominative	n		x	x	x	x	x
	genitive	g		x	x		x	x
	dative	d		x	x			x
	accusative	a		x	x			x
	vocative	v	x		x	x		
	locative	l		x	x			
	instrumental	i		x	x			x
l.s.	direct	r	x					
l.s.	oblique	o	x					
l.s.	partitive	1					x	
	illative	x					x	x
	inessive	2					x	x
	elative	e					x	x
...								
l.s.	temporalis	m						x
l.s.	causalis	c						x
l.s.	sublative	s						x
l.s.	delative	h						x
l.s.	sociative	q						x
l.s.	factive	y						x
l.s.	superessive	p						x
l.s.	distributive	u						x

Learning morphology: the case of the past tense of English verbs (with FOIDL)

- Examples in orthographic form:

```
past([s,l,e,e,p],[s,l,e,p,t])
```

- Background knowledge for FOIDL contained the predicate

```
split(Word,Prefix,Suffix), which works on nonempty lists
```

- An example decision list induced from 250 examples:

```
past([g,o],[w,e,n,t]) :- !.
```

```
past(A,B) :- split(A,C,[e,p]),split(B,C,[p,t]),!.
```

...

```
past(A,B) :- split(B,A,[d]), split(A,C,[e]),!.
```

```
past(A,B) :- split(B,A,[e,d]).
```

- Mooney and Califf (1995) report much higher accuracy on unseen cases as compared to a variety of propositional approaches

Learning first-order decision lists: FOIDL

- FOIDL (Mooney and Califf, 1995)
- Learns ordered lists of Prolog clauses,
a cut after each clause
- Learns from positive examples only
(makes output completeness assumption)
- Decision lists correspond to rules that use the Elsewhere Condition, which is well known in morphological theory
- They are thus a natural representation
for word-formation rules

Learning Slovene (nominal) inflections

The Slovene language has a rich system of inflections

Nouns in Slovene are lexically marked for **gender** (masculine, feminine or neuter)

They inflect for **number** (singular, plural or dual) and **case** (nominative, genitive, dative, accusative, locative, instrumental)

The paradigm of a noun consists of 18 morphologically distinct forms

Nouns can belong to different paradigm classes (declensions)

Alternations of inflected forms (stem and/or ending modifications) depend on morphophonological makeup, morphosyntactic properties, declension. Can also be idiosyncratic.

The paradigm of the noun golob (pigeon)

golob	=	Ncmsn	#singular nominative
goloba	golob	Ncmda	#dual accusative
goloba	golob	Ncmdn	#dual nominative
goloba	golob	Ncmsa	#singular accusative
goloba	golob	Ncmsg	#singular genitive
golobe	golob	Ncmpa	#plural accusative
golobi	golob	Ncmpi	#plural instrumental
golobi	golob	Ncmpn	#plural nominative
golobih	golob	Ncmdl	#dual locative
golobih	golob	Ncmpl	#plural locative
golobom	golob	Ncmpd	#plural dative
golobom	golob	Ncmsi	#singular instrumental
goloboma	golob	Ncmdd	#dual dative
goloboma	golob	Ncmdi	#dual instrumental
golobov	golob	Ncmdg	#dual genitive
golobov	golob	Ncmpg	#plural genitive
golobu	golob	Ncmsd	#singular dative
golobu	golob	Ncmsl	#singular locative

Ncm = Noun common masculine

Learning Slovene (nominal) inflections

Task

- Learn analysis and synthesis rules for Slovene (nominal) inflections
- Synthesis: *base form* \Rightarrow *oblique forms*
- Analysis: *oblique forms* \Rightarrow *base form*

Motivation

- Make it possible to analyse unknown words (not in lexicon). Analysis rules can infer the base form (and MSD) of such words.
- Compress the lexicon by storing rules + base forms only
Size(NewLex) approx. = 1/18 Size(OldLex) + Size of rules for A&S
- Make it easier to add new entries to the lexicon (only base)

The nominal paradigms dataset(s)

- Each MSD treated as a concept/predicate
`msd (Lemma , WordForm)`
- For synthesis, Lemma is input and WordForm output
- For analysis, WordForm is input and Lemma output
- A lexicon entry, e.g., `golob goloba Ncmsg`, gives rise to an example, e.g., `ncmsg (golob, goloba)`
- Common and proper nouns inflect in the same way, thus `Nc` and `Np` collapsed to `Nx`
- Orthographic representation of lemmas and word-forms used: `nxmsg ([g,o,l,o,b] , [g,o,l,o,b,a])`.

The nominal paradigms dataset(s)

- Syncretisms (word-forms always identical to some other word-forms).

Dual genitive = plural genitive, neuter accusative = neuter nominative

- Syncretisms omitted, leaving 37 concepts to learn
- The remaining MSDs and the corresponding dataset sizes are as follows

<code>nxmsn</code> = 2859	<code>nxfsn</code> = 2772	<code>nxnsn</code> = 1350
<code>nxmsg</code> = 2926	<code>nxfsg</code> = 2769	<code>nxnsg</code> = 1349
<code>nxmsd</code> = 2880	<code>nxfsd</code> = 2767	<code>nxnsd</code> = 1349
<code>nxmsa</code> = 2888	<code>nxfsa</code> = 2769	
<code>nxmsl</code> = 2885	<code>nxfsl</code> = 2767	<code>nxnsl</code> = 1349
<code>nxmsi</code> = 2886	<code>nxfsi</code> = 2773	<code>nxnsi</code> = 1349
<code>nxmpn</code> = 2777	<code>nxfpn</code> = 2585	<code>nxnnpn</code> = 1253
<code>nxmpg</code> = 2726	<code>nxfpg</code> = 2599	<code>nxnpg</code> = 1254
<code>nxmpd</code> = 2761	<code>nxfpd</code> = 2582	<code>nxnpd</code> = 1253
<code>nxmpa</code> = 2738	<code>nxfpa</code> = 2585	
<code>nxmpl</code> = 2768	<code>nxfpl</code> = 2582	<code>nxnpl</code> = 1254
<code>nxmpi</code> = 2763	<code>nxfpi</code> = 2582	<code>nxnpi</code> = 1254

Experimental setup for learning Slovene nominal paradigms

- Use the Multext East Lexicon
- For each of the 37 Slovene MSDs conduct two experiments, one for synthesis, the other for analysis
- Dataset sizes range from 1242 to 2926 examples
- For each experiment, 200 examples randomly selected from the dataset are used for training, while the remaining examples are used for testing

Summary of synthesis results

• **msd(+ Lemma , - WordForm)**

• Average accuracy = **91.4%**

$nxf = 97.8\%$ $nxn = 96.9\%$ $nxm = 80.5\%$

• Average number of rules = 16.4 (9.1 exceptions, 7.3 generalizations)

• Highest accuracy: $nxfsg = 99.2\%$ (4/1 – 4 rules of which 1 exception)

• Lowest accuracy: $nxmsa = 49.6\%$ (74/50)

Next lowest: $nxmpi = 76.6\%$ (35/20)

• Masculine singular accusative is syncretic, but the referred to rule is not constant

– If the noun is animate then $Nxmsa = Nxmsg$

– If the noun is inanimate then $Nxmsa = Nxmsn$

• Lexicon contains no information on animacy

An example set of rules for synthesis: nxfsg

Accuracy: 99.2%

4 rules (1 exception + 3 generalisations):

1. *prikazen* => *prikazni*

```
nxfsg([p,r,i,k,a,z,e,n],[p,r,i,k,a,z,n,i]).
```

2. *dajatev* => *dajatve*

```
nxfsg(A,B):-  
split(A,C,[v]),split(A,D,[e,v]),split(B,D,[v,e]).
```

3. *krava* => *krave*

```
nxfsg(A,B) :- split(A,C,[a]),split(B,C,[e]).
```

4. *prst* => *prsti*

```
nxfsg(A,B):-split(B,A,[i]).
```

Another set of rules for synthesis: nxmsg

Accuracy: 89.1%

27 rules (18 exception + 9 generalisations):

`nxmsg(A,B) :- split(A,C,[a]split(B,C,[a])).`

`nxmsg(A,B) :- split(A,C,[o]), split(B,C,[a]).`

-e- elision

`nxmsg(A,B) :- split(A,C,[z,e,m]), split(B,C,[z,m,a]).`

`nxmsg(A,B) :- split(A,C,[e,k]), split(B,C,[k,a]).`

`nxmsg(A,B) :- split(A,C,[e,c]), split(B,C,[c,a]).`

Stem lengthening by *-j-*

`nxmsg(A,B) :- split(B,A,[j,a]), split(A,C,[r]), split(A,[k],D).`

`nxmsg(A,B) :- split(B,A,[j,a]), split(A,C,[r]), split(A,[t],D).`

`nxmsg(A,B) :- split(B,A,[j,a]), split(A,C,[r]), split(A,D,[a,r]).`

`nxmsg(A,B) :- split(B,A,[a]).`

Summary of analysis results

- **msd(+ *WordForm* ,- *Lemma*)**
- Average accuracy = **91.5%**
nx~~f~~ = 94.8% nxn = 95.9% nx~~m~~ = 84.5%
- Average number of rules = 19.5 (10.5 exceptions, 9.1 generalizations)
- Highest accuracy: **nx~~n~~dd = 99.2% (5/2)**
- Lowest accuracy: **nx~~m~~dd = 82.1% (39/27)**

An example set of rules for analysis: nxfsg

Accuracy: 98.9%

6 rules (2 exceptions + 4 generalisations):

1. *prikazni* => *prikazen*

2. *ponve* => *ponev*

3. *dajatve* => *dajatev*

`nxfsg(A,B):-split(A,C,[v,e]),split(B,C,[e,v]),split(A,D,[a,t,v,e])`

4. *delitve* => *delitev*

`nxfsg(A,B):-split(A,C,[v,e]),split(B,C,[e,v]),split(A,D,[i,t,v,e]).`

5. *krava* => *krave*

`nxfsg(A,B) :- split(A,C,[e]),split(B,C,[a]).`

6. *prst* => *prsti*

`nxfsg(A,B):-split(A,B,[i]).`

Learning Slovene nominal inflections: Summary

- FOIDL (First-Order Induction of Decision Lists), shown to perform better than propositional systems on a similar problem, applied to learn nominal paradigms in Slovene
- Orthographic representation used
- For each MSD, 200 examples from lexicon taken as training examples
 - Rules learned for analysis/synthesis, tested on remaining entries
- Limited background knowledge used (splitting lists)
- Relatively good overall performance (average accuracy of 91.5%)
- Errors by the learned rules due to insufficient lexical information:
 - Orthography does not completely determine phonological alterations (e.g. schwa elision)
 - Morphosyntactic information missing (e.g. animacy)

Follow up work

- Uses CLOG instead of FOIDL to learn morphological rules
- Learning morphological analysis and synthesis rules for all Slovene MSDs
- Learning morphological analysis and synthesis rules for all MultextEast languages
- Learning POS tagging for Slovene
(with ILP and 4 other methods)
- Learning to lemmatize Slovene words

LEMMATIZATION

- The Task: Given wordform (but not MSD!), find lemma
- Motivation: Useful for lexical analysis
 - automated construction of lexica
 - information retrieval
 - machine translation
- One approach: lemma = stem
 - easy for English, but problems with inflections
 - user unfriendly
- Our approach: lemma = headword

LEMMATIZATION OF KNOWN AND UNKNOWN WORDS

- Given a large lexicon, known words can be lemmatized accurately, but ambiguously (*hotela* can be lemmatized to *hoteti* or *hotel*)
- Unambiguous lemmatization only possible if context taken into account (Part-Of-Speech=POS tagging used: *hoteti* is a Verb, *hotel* is a Noun)
- For unknown words, no lookup possible: rules/models needed
- To lemmatize unknown words in a given text
 - tag the given text with morphosyntactic tags
 - morphological analysis of the unknown words to find the lemmas

LEARNING TO LEMMATIZE UNKNOWN NOUNS, ADJECTIVES, AND VERBS

- Use existing annotated corpus to
- Learn a Part-Of-Speech tagger for a morphosyntactic tagset
(example tag: Ncmpi=Noun common masculine plural instrumental)
- Learn rules for morphological analysis of **open word classes**,
i.e., nouns, adjectives and verbs
(given morphosyntactic tag and wordform, derive lemma)
- Part of the corpus used for training, part for validation
- A separate testing set coming from a different corpus used

LEARNING MORPHOSYNTACTIC TAGGING

- Use the lexicon for training data
- Tagset of 1024 tags
(sentence boundary, 13 punctuation tags, 1010 morphosyntactic tags)
- Used the TnT (Brants, 2000) trigram tagger
- Also tried
 - Brill's Rule Based Tagger (RBT)
 - Ratnaparkhi's Maximum Entropy Tagger (MET)
 - Daelemans' Memory Based Tagger (MBT)

LEARNING MORPHOSYNTACTIC TAGGING

TnT constructs a table of n-grams (n=1,2,3)

...	
Vcps-sma	544
Vcip3s--n	82
Afpmsnn	17
Aopmsn	2
Ncmsn	12
Npmsn	1
Css	2
Afpnpa	1
Q	3

and a lexicon of wordforms

...					
juhe	2	Ncfsg	2		
julij	1	Npmsn	1		
julija	59	Npfsn	58	Ncmsa--n	1
julije	4	Npfsg	4		
juliji	10	Npfsd	10		
julijin	4	Aspmsa--n	2	Aspmsn	2
...					

THE TRAINING DATA

“1984” by George Orwell (Slovene translation) from MULTEXT-East project

- Lexicon for morphology, corpus for PoS tagging

- Inflection

	Slovene	English
Words	90,792	104,286
Forms	16,401	9,181
Lemmas	7,903	7,059
MSDs	1,010	134

- The lexical training set

PoS	Entries	WForms	Lemmas	MSDs
Noun	124,988	60,133	7,278	99
Adjective	306,746	63,764	4,551	279
Main Verb	110,295	77,533	3,682	43
All	542,029	194,142	15,479	421

THE TESTING DATA

IJS-ELAN Corpus

- Developed with the purpose of use in language engineering and for translation and terminology studies
- Composed of fifteen recent terminology-rich texts and their translations
- Contains 1 million words, about half in Slovene and half in English

- Size

	Slovene	English
Translation segments	31,900	31,900
Punctuation tokens	90,279	83,761
Word tokens	501,437	590,575
Word types	50,331	24,377
Lexical words	43,278	20,592

OVERALL EXPERIMENTAL SETUP

1. From the MULTEXT-East Lexicon (MEL)
for each MSD in the open word classes:
Learn rules for morphological analysis using CLOG
2. From the MULTEXT-East "1984" tagged corpus (MEC) :
Learn a tagger T0 using TnT
3. From IJS-ELAN untagged corpus (IEC)
take a small subset S0 (of cca 1000 words):
Evaluate performance of T0 on this sample (~ 70% – quite low)
4. From IEC take a subset S1 (of cca 5000 words),
manually tag and validate:
Learn a tagger T1 from MEC U S1 using TnT

5. Use a large backup lexicon (AML) that provides the ambiguity classes:

Lematize IEC using this lexicon and estimate the frequencies of MSDs within ambiguity classes using the tagged corpus MEC [S1]

6. From IEC take a subset S2 of (cca 5000 words), tag it with T1 + AML yielding IEC-T, manually validate:

This gives an estimate of tagging accuracy

7. Take the tagged and lematized IEC-T, extract all open class inflecting word tokens which posses a lemma (were in the AML lexicon) yielding the set AK; those that do not posses a lemma go to LU

8. Test the analyzer on AK

9. Test the lemmatiser (consisting of the tagger+analyzer) on LU

TAGGING RESULTS ON THE IJS-ELAN CORPUS

	All	Errors	Accuracy
Nouns	1276	133	89.6%
Adjectives	499	77	84.6%
Main Verbs	505	17	96.6%
Open	2280	227	90.0%
Words	3820	318	91.7%
Tokens	4454	318	92.9%

MORPHOLOGICAL ANALYSIS RESULTS ON THE TESTING DATASET (IJS-ELAN)

PoS	Entries	Error	Accuracy	Baseline
Noun	4,834	85	98.2%	31.9%
Adjective	4,764	50	98.9%	8.9%
Main Verb	588	10	98.3%	11.9%
All	10,186	145	98.6%	20.0%

LEMMATIZATION RESULTS ON THE TESTING DATASET (IJS-ELAN)

PoS	Entries	Error	Accuracy
Noun	405	36	91.1%
Adjective	308	16	94.8%
Main Verb	50	9	82.0%
All	763	61	92.0%

- Accuracy of tagging for unknown nouns/adjectives/verbs 90.0%
- Accuracy of analysis for unknown nouns and adjectives 98.6%
- Accuracy of lemmatization for unknown nouns and adjectives 92.0%
- Main source of error is tagger error, which doesn't always hurt analysis (syncretism)
- Most serious error is when tagger gives a wrong wordclass

Learning Lemmatization: Summary

CONCLUSIONS AND FURTHER WORK

- Learned to lemmatize unknown nouns and adjectives by learning morphosyntactic tagging and morphological analysis

- Accuracy of 92% on new text

- High above baseline accuracy

If we say lemma=wordform, we get accuracy of approximately 40%

- Comparison with other approaches to lemmatizing unknown Slovene words

- Learn better tagger

- Learn from larger corpus/corpora

MultextEast for Macedonian

- On-going work
- Bilateral project SI-MK:
Gathering, Annotation and Analysis of
Macedonian/Slovenian Language Resources
- PIs: Katerina Zdravkova, Saso Dzeroski
- Creating the MK version of the “1984”
corpus, as well as a corresponding lexicon

MultextEast for Macedonian

- Creation of the “1984” corpus
 - Scanning of the cyrillic version of the novel
 - OCR
 - Error correction (spell-checking & manual)
 - Tokenization
 - Conversion to XML (TEI compliant)
 - Alignment (with the English “1984” original)
 - BSc Thesis of Viktor Vojnovski

Multext East for Macedonian

- Morphosyntactic specifications
- Macedonian nouns have 5 attributes:
 - type (common, proper)
 - gender (masculine, feminine, neuter)
 - number (singular, plural, count)
 - case (nominative, vocative, oblique)
 - definiteness (no, yes, close, distant)
- Manual annotation
 - Complete for nouns
 - Only PoS for other word categories

MultextEast for Macedonian

Applying Machine Learning

- Learning morphological analysis and synthesis (BSc thesis Aneta Ivanovska)
- Learning PoS tagging (with incomplete tagset/ full tags only for nouns/ PoS only for the rest; BSc thesis Viktor Vojnovski)
- Example: Analysis rules for Feminine nouns, plural, nominative, nondefinite

Exceptions:

raspravii -> rasprava

strui -> struja

race -> raka

noze -> noga

boi -> boja

Rules:

*sti -> *st

*ii -> *ija

id*i -> id*ja

*i -> *a

Talk outline

- Language technologies and linguistics
- Language resources
- The Multext-East resources
 - Learning morphological analysis/synthesis
 - Learning PoS tagging
 - Lemmatization
- The Prague Dependency Treebank
 - Learning to assign tectogrammatical functors

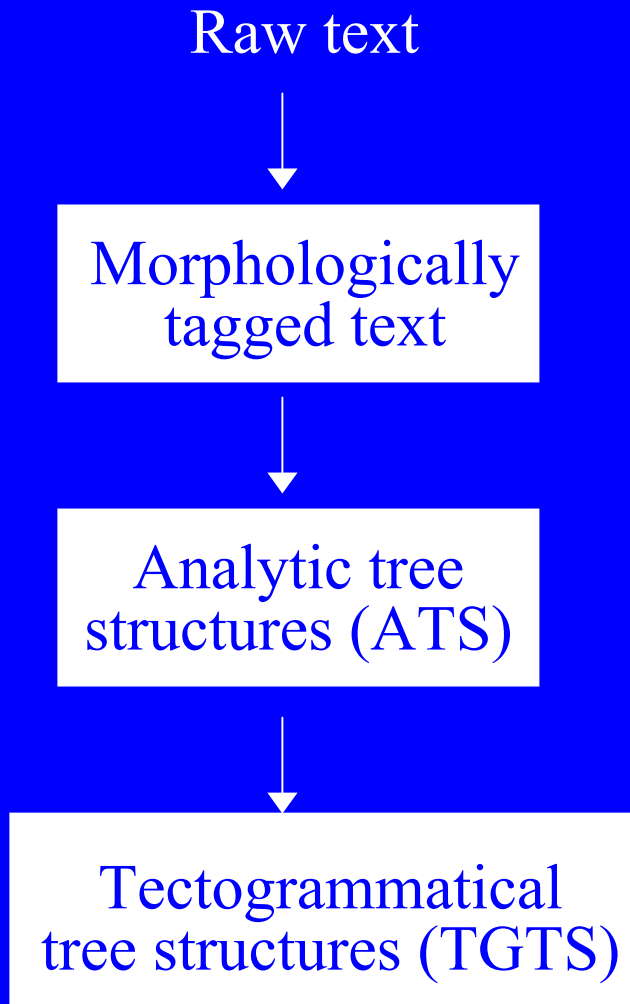
Prague Dependency Treebank (PDT)

- Long-term project aimed at a complex annotation of a part of the Czech National Corpus with rich annotation scheme
- Institute of Formal and Applied Linguistics
 - Established in 1990 at the Faculty of Mathematics and Physics, Charles University, Prague
 - Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall
 - <http://ufal.mff.cuni.cz>

Prague Dependency Treebank

- Inspiration:
 - The Penn Treebank (the most widely used syntactically annotated corpus of English)
- Motivation:
 - The treebank can be used for further linguistic research
 - More accurate results can be obtained (on a number of tasks) when using annotated corpora than when using raw texts
- PDT reaches representations suitable as input for semantic interpretation, unlike most other annotations

Layered structure of PDT



- Morphological level
 - Full morphological tagging (word forms, lemmas, mor. tags)
- Analytical level
 - Surface syntax
 - Syntactic annotation using dependency syntax (captures analytical functions such as subject, object,...)
- Tectogrammatical level
 - Level of linguistic meaning (tectogrammatical functions such as actor, patient,...)

The Analytical Level

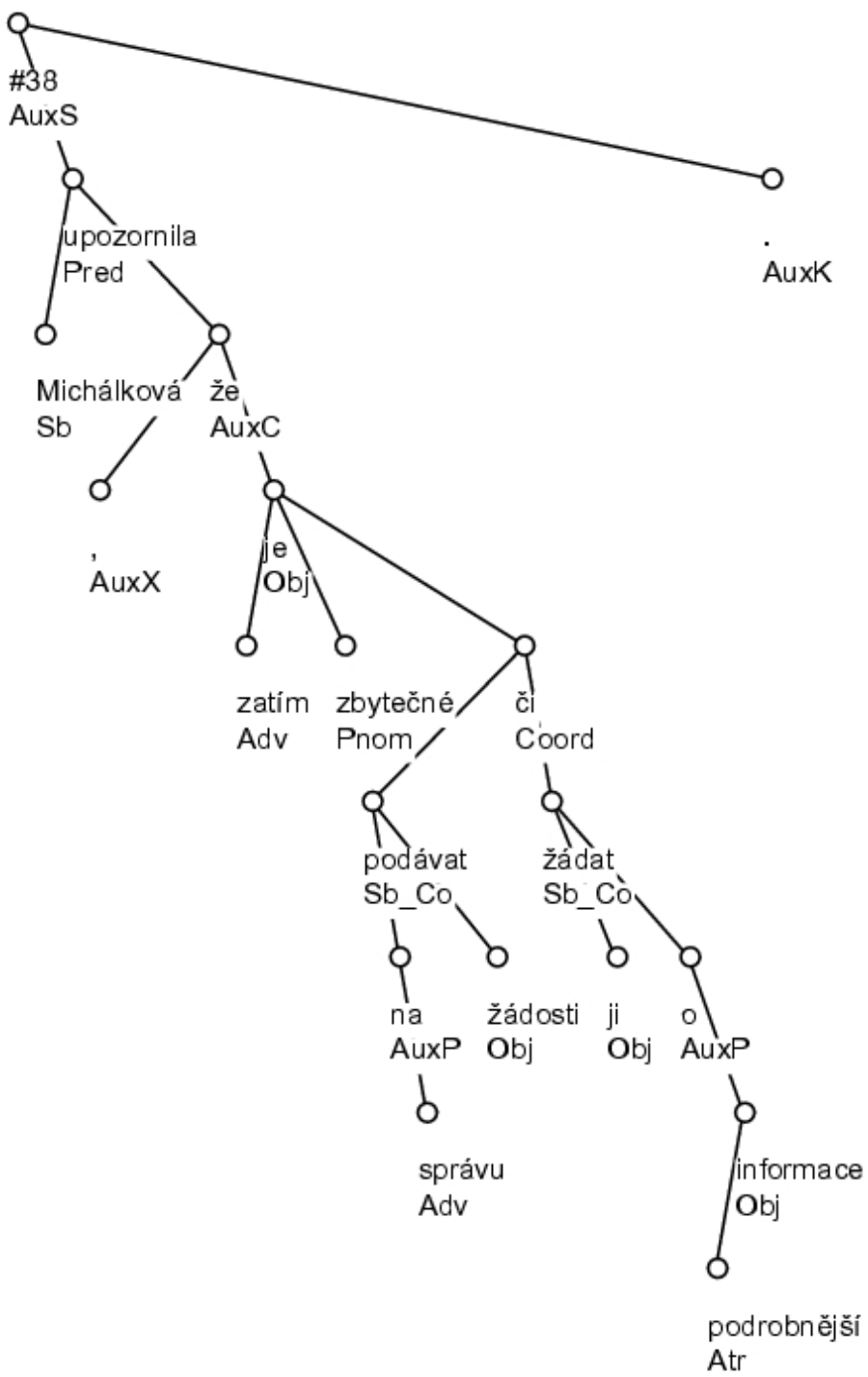
- The dependency structure chosen to represent the syntactic relations within the sentence
- Output of the analytical level: analytical tree structure
 - Oriented, acyclic graph with one entry node
 - Every word form and punctuation mark is a node
 - The nodes are annotated by attribute-value pairs
- New attribute: analytical function
 - Determines the relation between the dependent node and its governing nodes
 - Values: Sb, Obj, Adv, Atr,....

The Tectogrammatical Level

- Based on the framework of the Functional Generative Description as developed by Petr Sgall
- In comparison to the ATs, the tectogrammatical tree structures (TGTSs) have the following characteristics:
 - Only autosemantic words have an own node, function words (conjunctions, prepositions) are attached as indices to the autosemantic words to which they belong
 - Nodes are added in case of clearly specified deletions on the surface level
 - Analytical functions are substituted by tectogrammatical functions (functors), such as Actor, Patient, Addressee,...

Functors

- Tectogrammatical counterparts of analytical functions
- About 60 functors
 - Arguments (or theta roles) and adjuncts
 - Actants (Actor, Patient, Addressee, Origin, Effect)
 - Free modifiers (LOC, RSTR, TWHEN, THL,...)
- Provide more detailed information about the relation to the governing node than the analytical function



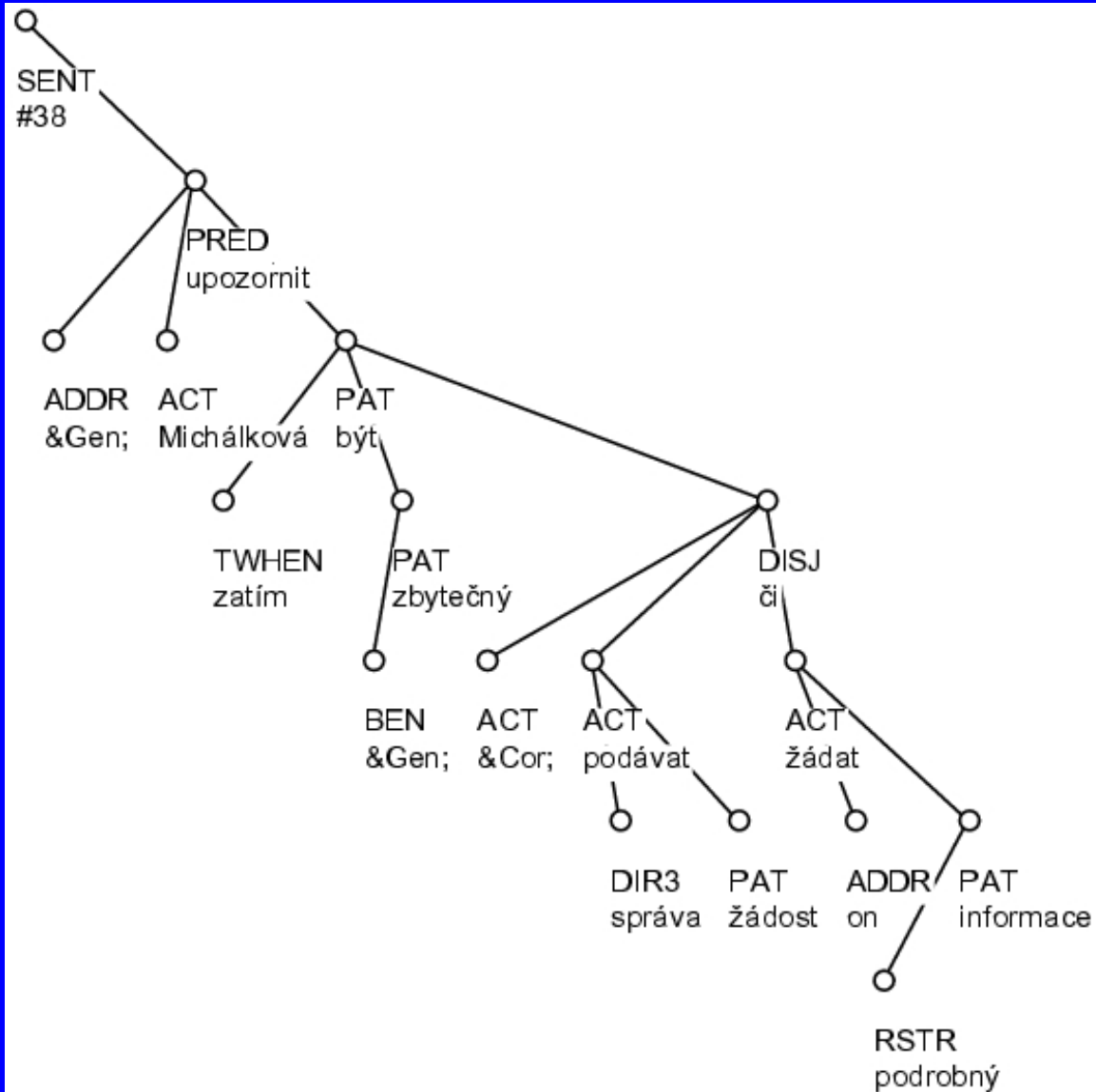
AN EXAMPLE ATS:

Michalkova upozornila, že zatím je zbytečné podávat na správu žádosti či žádat ji o podrobnější informace.

Literal translation:

Michalkova pointed-out that meanwhile is superfluous to-submit to administration requests or to-ask it for more-detailed information.

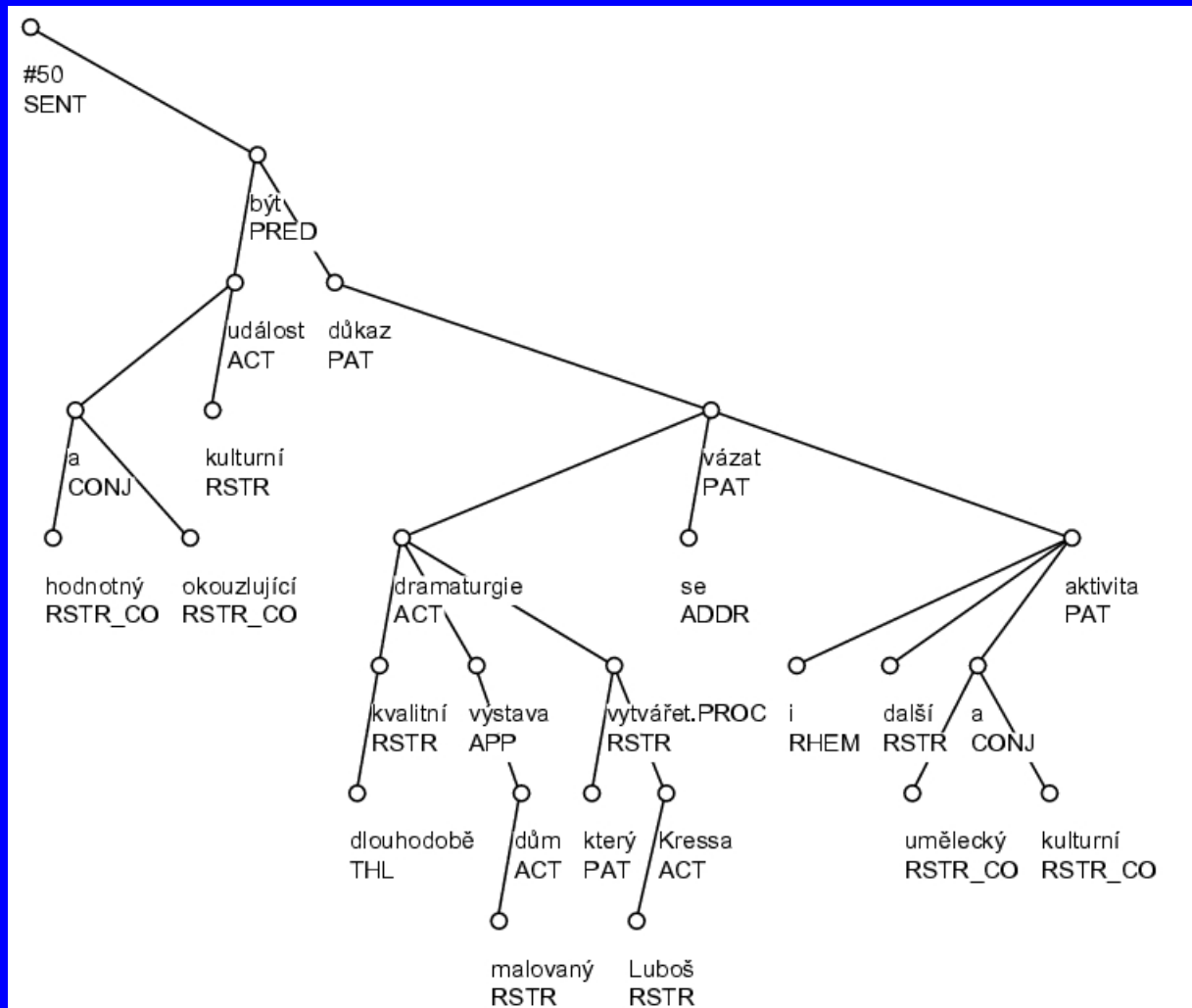
AN EXAMPLE TGTS FOR THE SENTENCE: “M. pointed out that for the time being it was superfluous to submit requests to the administration, or to ask it for a more detailed information.”



Literal translation:
 Michalkova pointed-out
 that meanwhile
 is superfluous to-submit
 to administration requests
 or to-ask it
 for more-detailed information.

AN EXAMPLE TGTS FOR THE SENTENCE:

“The valuable and fascinating cultural event documents that the long-term high-quality strategy of the Painted House exhibitions, established by L. K., attracts further activities in the domains of art and culture.”



Some TG Functors

ACMP (accompagnement): mothers with **children**

ACT (actor): **Peter** read a letter.

ADDR (addressee): Peter gave **Mary** a book.

ADVS (adversative): He came there, **but** didn't stay long.

AIM (aim): He came there to **look** for Jane.

APP (appuerenance, i.e., possession in a broader sense): **John's** desk

APPS (apposition): Charles the Fourth, (i.e.) **the Emperor**

ATT (attitude): They were here **willingly**.

BEN (benefactive): She made this for her **children**.

CAUS (cause): She did so since they **wanted** it.

COMPL (complement): They painted the wall **blue**.

COND (condition): If they **come** here, we'll be glad.

CONJ (conjunction): Jim **and** Jack

CPR (comparison): **taller** than Jack

CRIT (criterion): According to **Jim**, it was rainng there.

Some more TG Functors

ID (entity): the river **Thames**

LOC (locative): in **Italy**

MANN (manner): They did it **quickly**.

MAT (material): a bottle of **milk**

MEANS (means): He wrote it by **hand**.

MOD (mod): He **certainly** has done it.

PAR (parentheses): He has, as we **know**, done it yesterday.

PAT (patient): I saw **him**.

PHR (phraseme): in no **way**, grammar **school**

PREC (preceding, particle referring to context): **therefore, however**

PRED (predicate): I **saw** him.

REG (regard): with regard to **George**

RHEM (rhematizer, focus sensitive particle): **only, even, also**

RSTR (restrictive adjunct): a **rich** family

THL (temporal-how-long): We were there for three **weeks**.

THO (temporal-how-often) We were there very **often**.

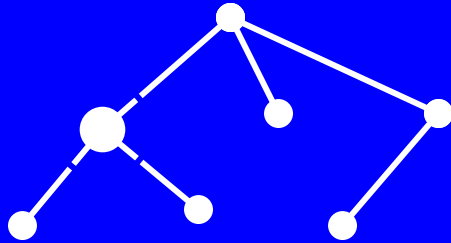
TWHEN (temporal-when): We were there at **noon**.

Automatic Functor Assignment

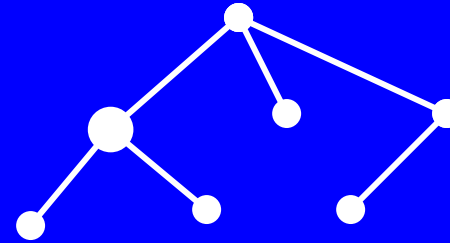
- Motivation: Currently annotation done by humans, consumes huge amounts of time of linguistic experts
- Overall goal: Given an ATS, generate a TGTS
- Specific task: Given a node in an ATS, assign a tectogrammatical functor
- Approach: Use sentences with existing manually derived ATSs and TGTSs to **learn** how to assign tectogrammatical functors
- More specifically, use machine learning to learn rules for assigning tectogrammatical functors

What context of a node to take into account for AFA purposes?

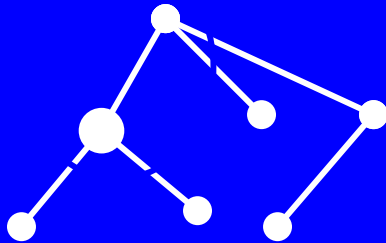
a) only node U



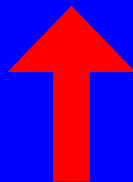
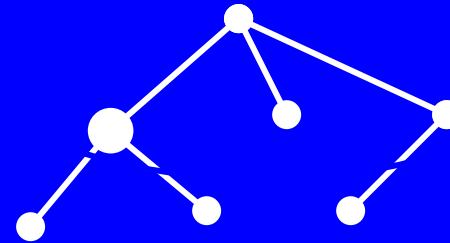
b) whole tree



c) node U and its parent



d) node U and its siblings



The attributes

- **Lexical attributes:** lemmas of both G and D nodes, and the lemma of a preposition / subordinating conjunction that binds both nodes,
- **Morphological attributes:** POS, subPOS, morphological voice, morphologic case,
- **Analytical attributes:** the analytical functors of G/D
- **Topological attributes:** number of children (directly depending nodes) of both nodes in the TGTS
- **Ontological attributes:** semantic position of the node lemma within the EuroWordNet Top Ontology

AFA - Take 1 (2000):

The attributes and the class

Given

Governing node

- Word form
- Lemma
- Full morphological tag
- Part of speech (POS)
(extracted from above)
- Analytical function
from ATS

Dependent node

- Word form
- Lemma
- Full morphological tag
- POS and case
(extracted from above)
- Analytical function

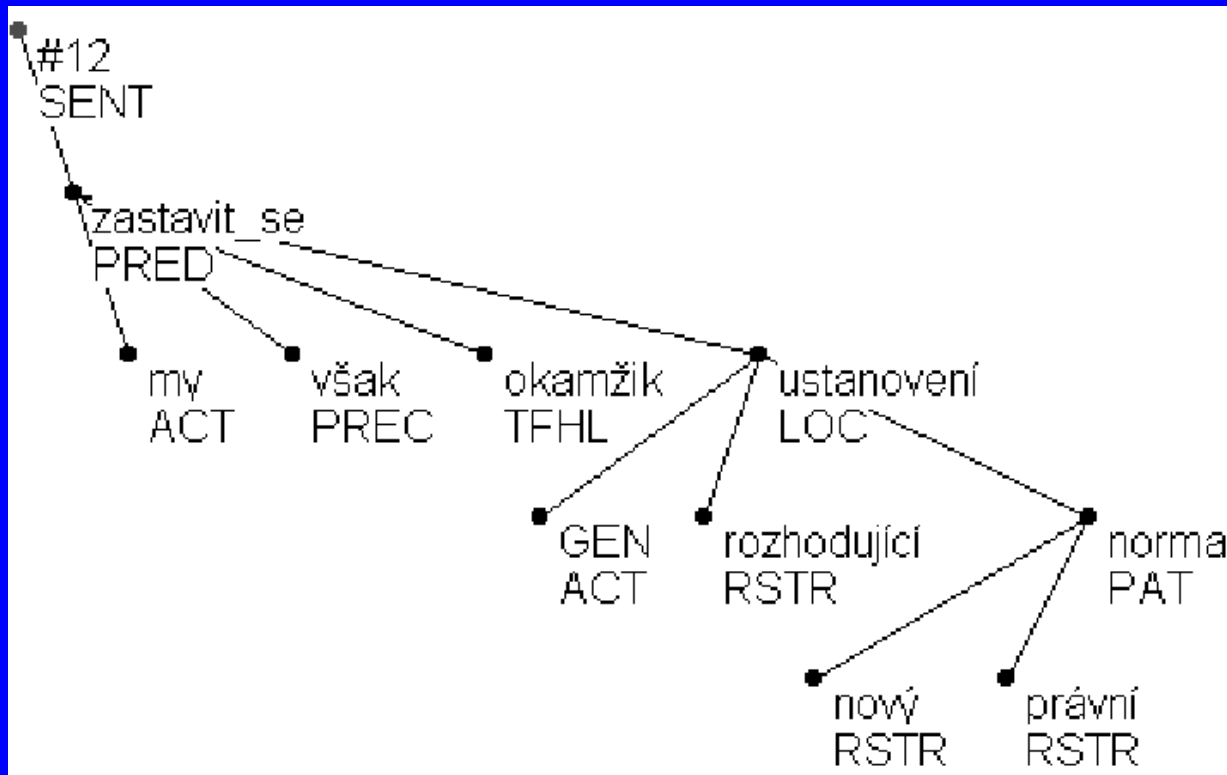
Conj. or preposition

between G and D node

Predict: Functor of the dependent node

Training examples

zastavme :zastavit1 :vmp1a:v:pred:okamz_ik :okamz_ik :nis4a :n:4:na:adv:tfhl
zastavme :zastavit1 :vmp1a:v:pred:ustanoveni_ :ustanoveni_ :nns2a :n:2:u :adv :loc
normy :norma :nfs2a :n:atr :nove_ :novy_ :afs21a :a:0: :atr :rstr
normy :norma :nfs2a :n:atr :pra_vni_ :pra_vni_ :afs21a:a:0: :atr :rstr
ustanoveni_ :ustanoveni_ :nns2a :n:adv:normy :norma :nfs2a :n:2: :atr :pat



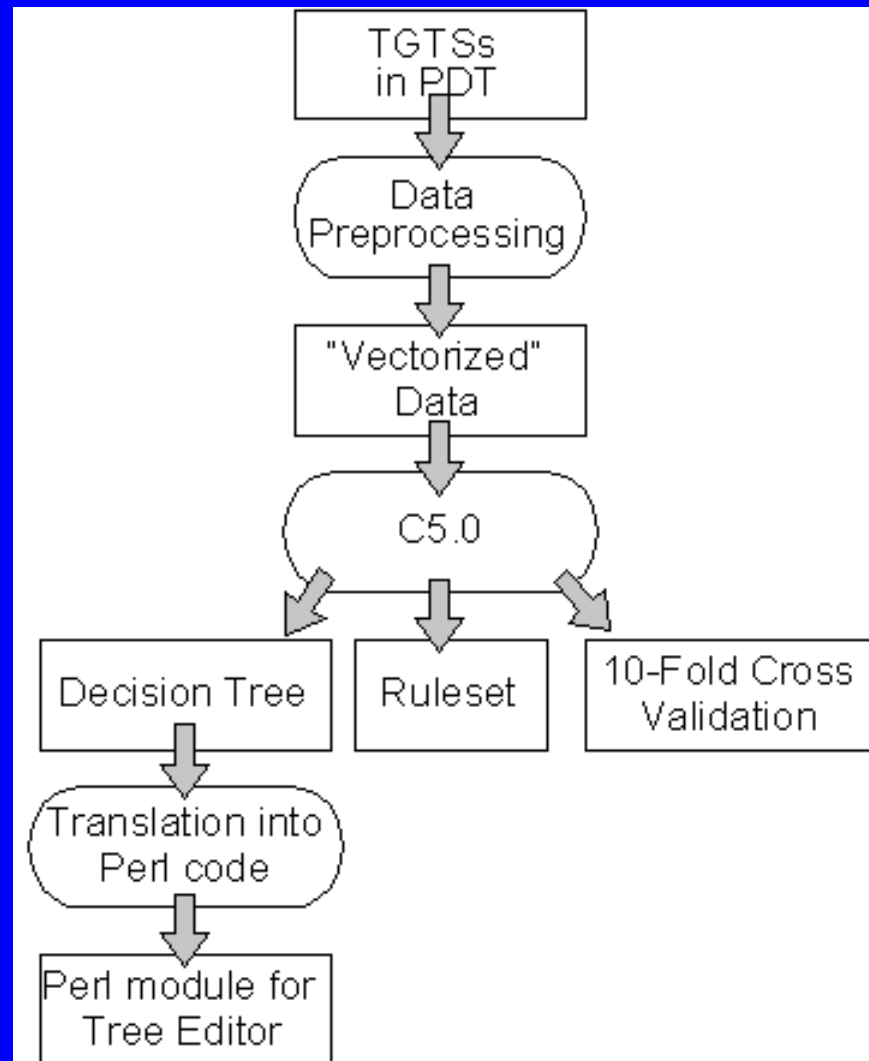
AFA - Take 2 (2002)

- In Take 1, ML and hand-crafted rules used
- Lesson from Take 1: Annotators want high recall, even at the cost of lower precision
- Use machine learning only
- More training data/annotated sentences (1536 sentences; 27463 nodes in total)
- Use a larger set of attributes
 - Topological (number of children of G/D nodes)
 - Ontological (WordNet)
- We use the ML method of decision trees (C5.0)

Ontological attributes

- Semantic concepts (63) of Top Ontology in EWN (e.g., Place, Time, Human, Group, Living, ...)
- For each English synset, a subset of these is linked
- Inter Lingual Index – Czech lemma -> English synset -> subset of semantic concepts
- 63 binary attributes: positive/negative relation of Czech lemma to the respective concept TOEWN

Methodology



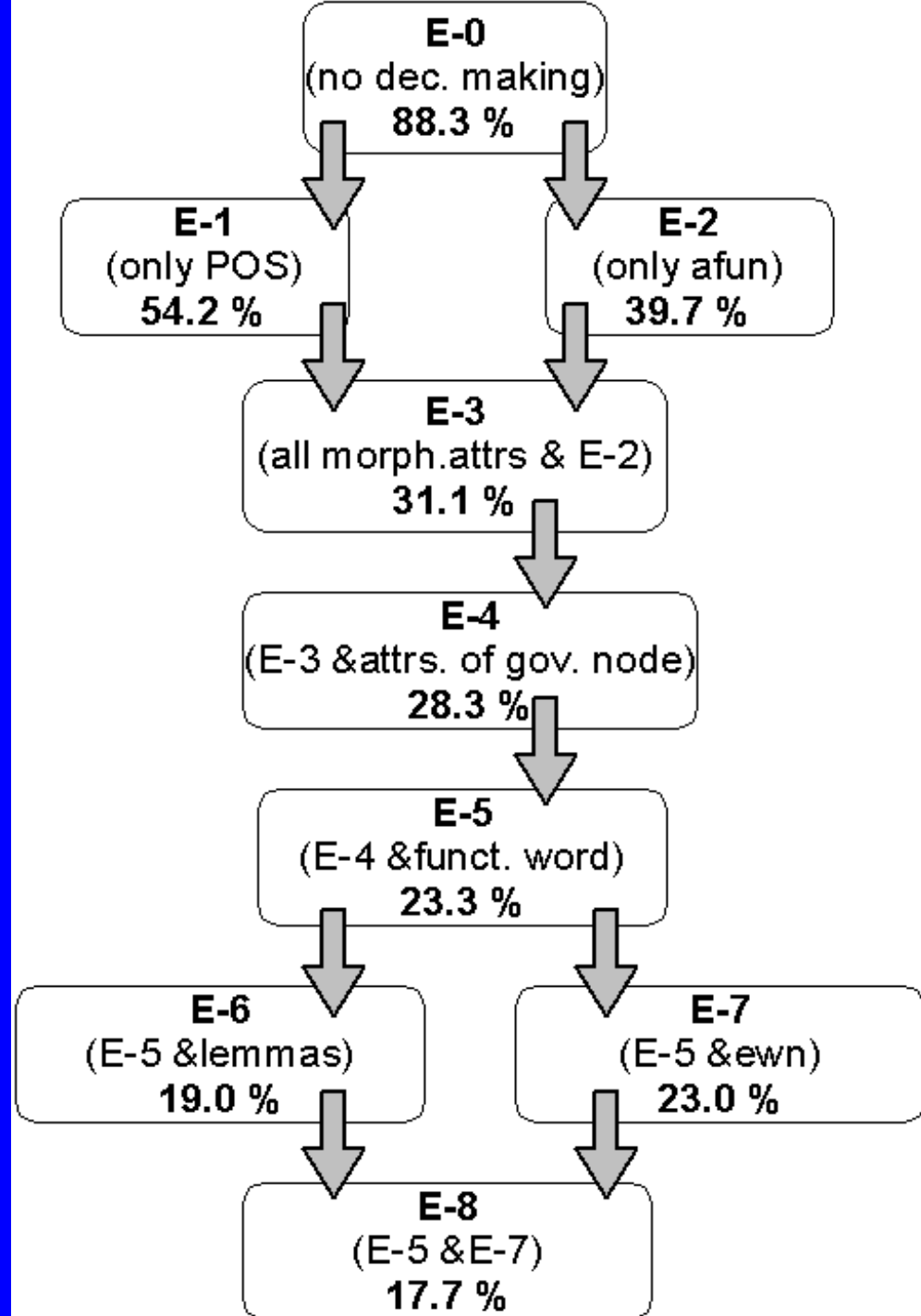
Methodology

- Evaluation of accuracy by 10-fold cross-validation
- Rules to illustrate the learned concepts
- Trees translated to Perl code included in TrEd – a tool that annotators use

Different sets of attributes

- E-0 (empty)
- E1 – Only POS; E2 – Only Analytical function
- E3 – All morphological atts & E-2
- E4 – E3 & Attributes of governing node
- E5 – E4 & funct. Words (preps./conjs.)
- E6 – E5 & lemmas; E7 – E5 & EWN
- E8 – E6 & E7

AFA performance



Example rules (1)

E-1) Input attributes: part of speech of node N

Error rate: 54.2%

Decision tree size: 13 leaves

Sample from the ruleset:

Rule 7: (3584/461, lift 4.0)

d_pos = A

-> class RSTR [0.871]

Possible interpretation of the rule: An adjective usually is the restrictive adjunct.

Example rules (2)

E-2) Input attributes: analytical function of given node.

Error rate: 39.7%

Decision tree size: 45 leaves

Sample from the ruleset:

```
Rule 21: (2244/323, lift 5.5)
         d_afun = Sb
         -> class ACT [0.856]
```

Interpretation: The subject of a sentence usually becomes its actor.

Example rules (3)

E-3) Input attributes: morphological attributes and analytical function of given node.

Error rate: 31.1%

Decision tree size: 416 leaves

Sample from the ruleset:

```
Rule 213: (251/130, lift 29.2)
  d_case = 3
  d_afun = Obj
  -> class ADDR [0.482]
```

Interpretation: An object in dative becomes addressee.

E-4) Input attributes: morphological attributes and analytical function of given node and of its autosemantic governor G

Error rate: 28.3%

Decision tree size: 1785 leaves

Rule 388: (16/4, lift 4.7)

g_voice = P

d_case = 7

d_afun = Obj

-> class ACT [0.722]

Rule 665: (137/14, lift 5.9)

g_voice = P

d_afun = Sb

-> class PAT [0.892]

Interpretation: The subject in a clause in passive voice becomes patient, the actor is expressed by instrumental (Compare with the rule in E-2).

E-5) Input attributes: Same attributes as in E-4, but lemmas of functional word (prepositions, conjunctions) were added.

Error rate: 23.3%

Decision tree size: 1716 leaves

Sample from the ruleset:

Rule 11: (63/16, lift 108.0)

d_afun = Adv

preposition = s

-> class ACMP [0.738]

Rule 174: (16, lift 231.1)

d_afun = Adv

subord_conj = protože

-> class CAUS [0.944]

Rule 412: (34/6, lift 368.7)

coord_conj = nebo

-> class DISJ [0.806]

Interpretations: (i) A node connected via preposition *s* ('with') represents accompaniment. (ii) A clause connected via subordinating conjunction *protože* ('because') relates to causality. (iii) A coordination node with lemma *nebo* ('or') expresses disjunction.

Example
rules (5)

E-6) Input attributes: same attributes as in E-5, but lemmas of both nodes were added.

Error rate: 19,0%

Decision tree size: 5037 leaves

Sample from the ruleset:

```
Rule 511: (40, lift 28.6)
  d_lemma = rok
  preposition = v
  -> class TWHEN [0.976]
```

```
Rule 1031: (6/3, lift 3.2)
  g_lemma = činnost
  d_pos = N
  preposition = empty
  -> class ACT [0.500]
```

```
Rule 617: (11, lift 736.0)
  d_lemma = dosud
  -> class TTILL [0.923]
```

```
Rule 1397: (6, lift 71.6)
  g_lemma = patřit
  preposition = mezi
  -> class DIR3 [0.875]
```

Interpretation: (i) *v roce* ('in year') is temporal modifier. (ii) A noun directly depending on noun *činnost* ('activity') is probably actor of the activity. (iii) *dosud* ('still', 'untill now') is a temporal modifier (TTILL - 'time till ...'). (iv)

Example rules (6)

E-7) Input attributes: morphological attributes and analytical function of node N

Error rate: 23.0%

Decision tree size: 1873 leaves

Sample from the ruleset:

```
Rule 237: (115, lift 29.0)
  d_afun = Adv
  preposition = v
  d_ewn_time = yes
  -> class TWHEN [0.991]
```

Interpretation: An adverbial formed by a noun that has (according to EuroWordNet) something to do with time and that is connected via preposition *v* ('in'), is a temporal modifier of type TWHEN.

Example rules (E8)

E-8) Input attributes: union of attributes from E-6 and E7

Error rate: 17.7%

Decision tree size: 4445 leaves

Sample from the ruleset:

Rule 70: (4, lift 132.9)

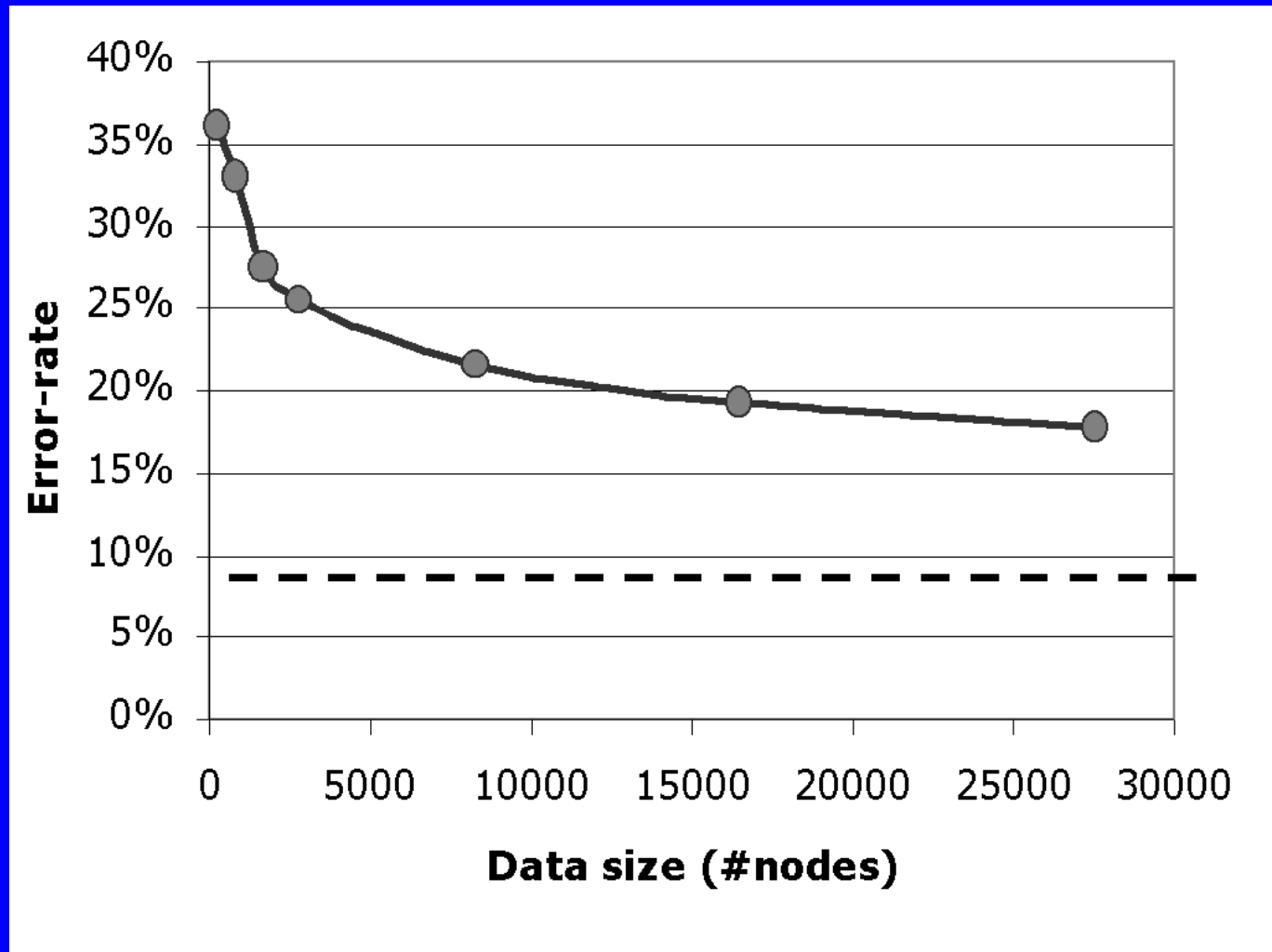
g_lemma = množství

d_ewn_origin = yes

-> class MAT [0.833]

Interpretation: If an item depends on noun *množství* ('amount') and it is related to concept Origin in EuroWordNet, then it has the functor MAT (material, e.g. amount of wood).

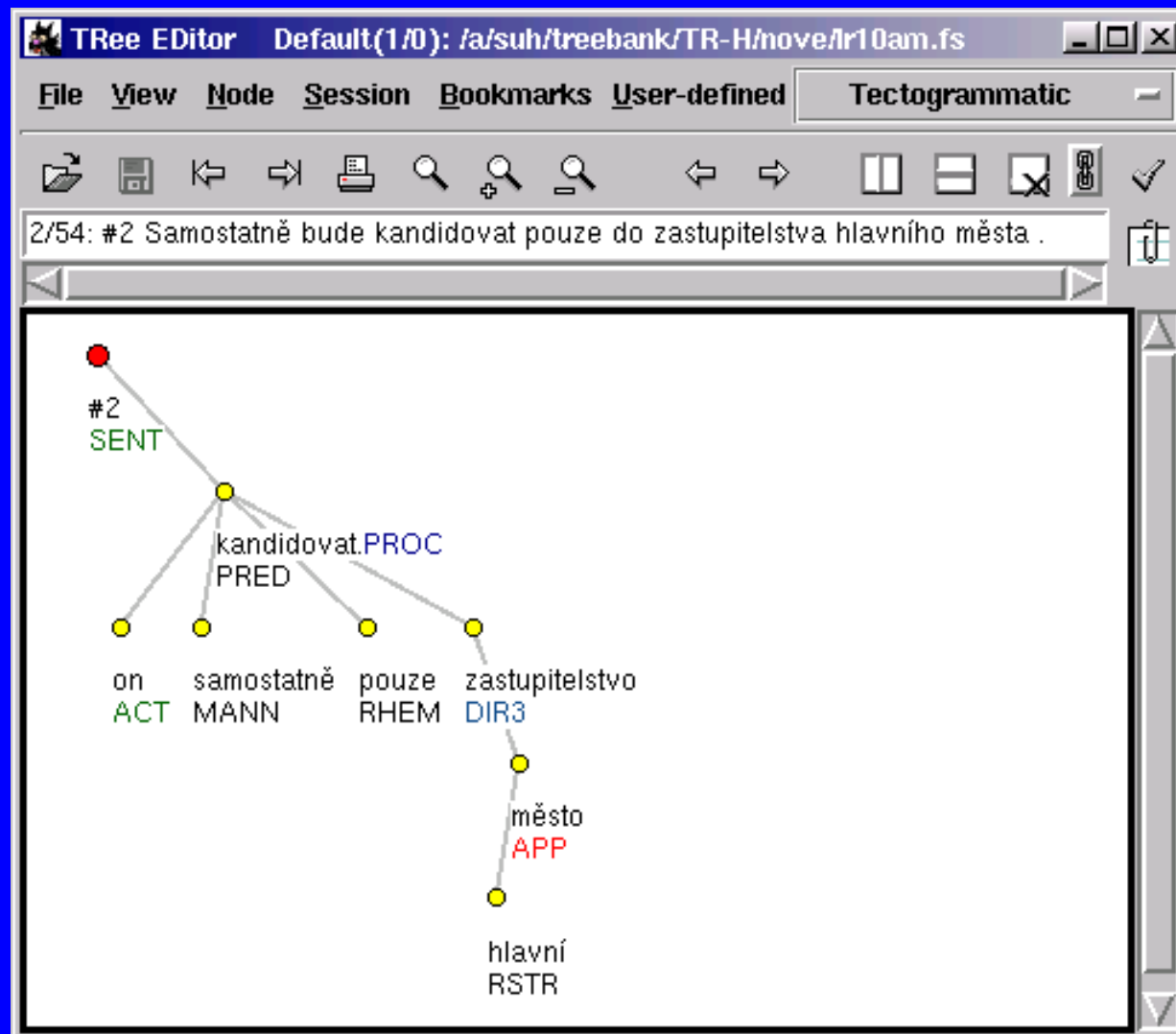
Learning curve (for E-8)



Using the learned AFA trees

- PDT Annotators use TrEd editor
- Learned trees transformed into Perl
- A keyboard shortcut defined in TrEd which executes the decision tree for each node of the TGT and assigns functors
- Color coding of factors based on confidence
 - Black: over 90%
 - Red: less than 60%
 - Blue: otherwise

Using the learned AFA trees in TrEd



Annotators response

- Six annotators
- All agree: The use of AFA significantly increases the speed of annotation (twice as long without it)
- All annotators prefer to have as many assigned functors as possible
- They do not use the colors (even though red nodes are corrected in 75% on unseen data)
- Found some systematic errors bade by AFA – suggested the use of topological attributes

PDT - Conclusions

- ML very helpful for annotating PDT, even though
- PDTs very close to the semantics of natural language
- Faster annotation
- Very accurate annotation
 - Automatically assigned functors corrected in 20 % of the cases
 - Human annotators disagree in more than 10% of the cases
 - Very close to what is possible to achieve through learning

Further work - SDT

- Slovene Dependency Treebank
- Morphological analysis (done)
- Part-Of-Speech tagging (done)
- Parsing/grammar (only a rough draft)
- Annotation of sentences
from Orwell's 1984 (in progress)

Summary

- (Annotated) language resources are very important
- We can use them to evaluate language tools
- And also create language tools by
- Using machine learning
- This for different levels of linguistic analysis, depending on the annotation of the resources

Further work

- Create language resources and tools for Slovenian and Macedonian
 - Corpora, treebanks
 - Dependency (ATs/TGTs) for SI/MK
 - Parsers for SI/MK
- Machine learning tools for this
 - Active learning
- Domain knowledge

Credits

- Tomaz Erjavec
- Jakub Zavrel
- Suresh Mannadhar, James Cussens
- Zdenek Zabokrtsky, Petr Sgall
- Aneta Ivanovska, Viktor Vojnovski
- Katerina Zdravkova