Linguistic Data Retrievable from a Treebank

Verginica Barbu Mititelu, Elena Irimia
Romanian Academy
Research Institute for Artificial Intelligence

Outline

1. The treebank
    1. The corpus
    2. Its processing and annotation
    3. Syntactic annotation
        1. Principles
        2. Relations
        3. Example
2. Data retrievable from the treebank
    1. Statistical data with linguistic relevance
    2. Data retrievable through querying
3. Conclusions

Outline

1. The treebank
    1. **The corpus**
    2. Its processing and annotation
    3. Syntactic annotation
        1. Principles
        2. Relations
        3. Example
2. Data retrievable from the treebank
    1. Statistical data with linguistic relevance
    2. Data retrievable through querying
3. Conclusions

1. Corpus structure

	Wiki	Acad	News	Biblio	EMEA	Frame Net	JRC	Lit	Medical	Misc	TOTAL
Sents	611	950	933	877	933	1092	1606	**1819**	277	424	**9522**
Tokens	14048	19991	23356	16876	19890	25654	**48295**	37308	7764	7959	**221141**
Length	23	21	25	19	21	23	**30**	21	28	19	**23**

Outline

1. The treebank
    1. The corpus
    2. **Its processing and annotation**
    3. Syntactic annotation
        1. Principles
        2. Relations
        3. Example
2. Data retrievable from the treebank
    1. Statistical data with linguistic relevance
    2. Data retrievable through querying
3. Conclusions

Corpus processing and annotation

- tokenisation
- lemmatisation
- morphologic analysis

} with TTL tool

Outline

1. The treebank
    1. The corpus
    2. Its processing and annotation
    3. **Syntactic annotation**
        1. Principles
        2. Relations
        3. Example
2. Data retrievable from the treebank
    1. Statistical data with linguistic relevance
    2. Data retrievable through querying
3. Conclusions

Syntactic annotation

- Dependency framework:
  - each sentence is analyzed as a tree (i.e., a directed acyclic graph)
  - Its nodes are the words and punctuation in the sentence
  - the edges are relations established between two nodes
  - All relations are hierarchical: the higher node in a relation is the head and the lower one is its dependent.
  - The only node without a head is the root.
  - Any head can have one or more dependents, or even none in the case of tree leaves.

Universal Dependency Principles

- the treatment of function words as dependents
- A flat structure (with the first occurring element as the head and all the others as its dependents) for coordination, multiword expressions, names, foreign, etc.
- Active and passive subjects and auxiliaries are marked distinctly.
- The clausal realisation of syntactic functions is marked distinctly from their lexical realisations.

Invetory of relations

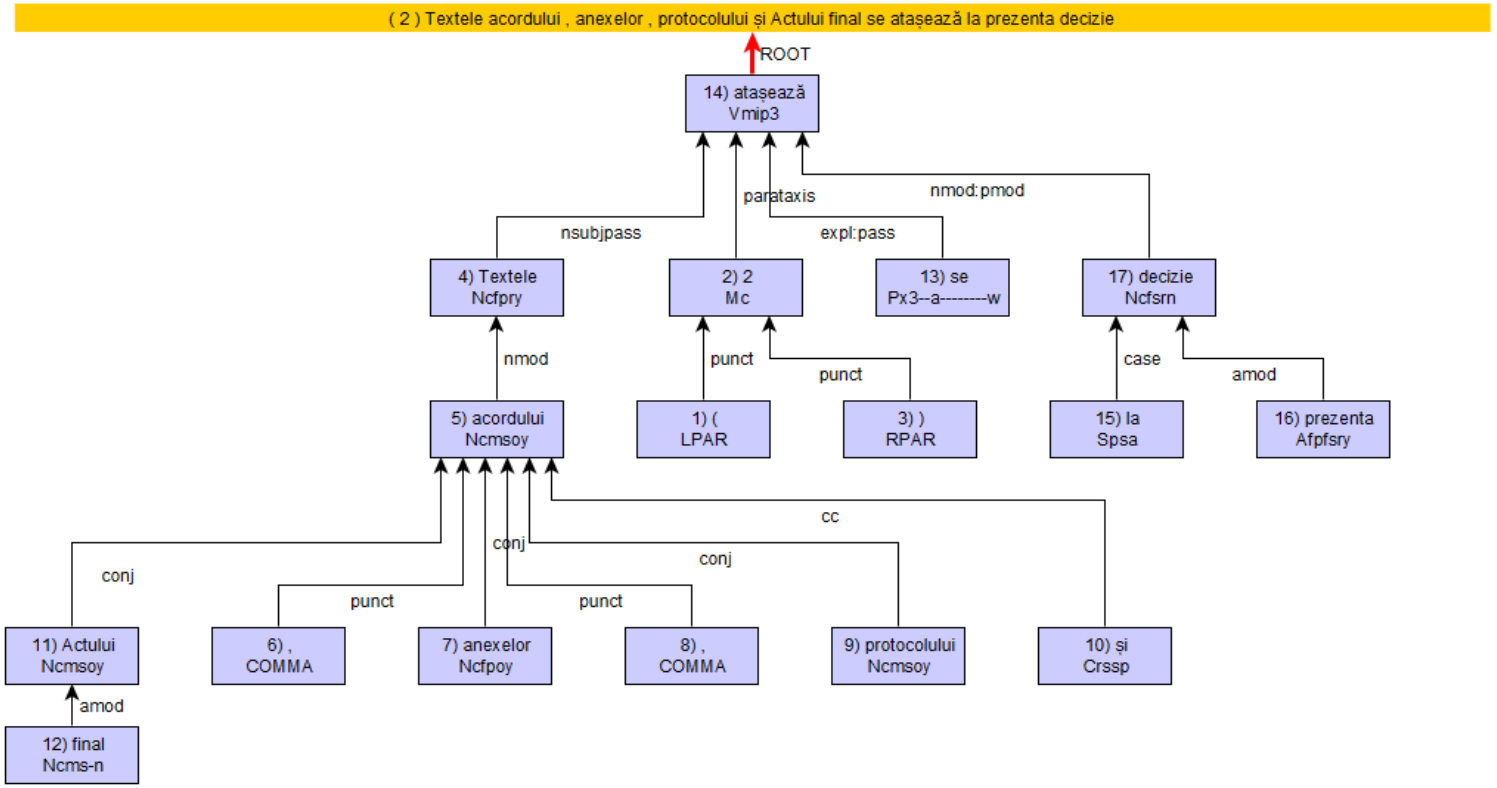
Core dependents of clausal predicates			Non-core dependents of clausal predicates			Special clausal dependents		
Nominal dep	**Predicate dep**		**Nominal dep**	**Predicate dep**	**Modifier word**	**Nominal dep**	**Auxiliary**	**Other**
nsubj	csubj		nmod	advcl	advmod	vocative	aux	mark
nsubjpass	csubjpass		***↳nmod:pmod***	***↳advcl:tcl***	***↳advmod:tmod***	discourse	auxpass	punct
dobj	ccomp	xcomp	**↳nmod:tmod**		neg	expl	cop	
iobj	***↳ccomp:pmod***		**↳nmod:agent**			**↳expl:pv**		
						↳expl:pass		
						↳expl:impers		
						↳expl:poss		
Noun dependents			**Compounding and unanalyzed**			**Coordination**		
Nominal dep	**Predicate dep**	**Modifier word**	compound	mwe		conj	cc	punct
nummod	acl	amod	name	foreign	goeswith		**↳cc:preconj**	
appos		det						
nmod		neg						
Case-marking, prepositions, possessive			**Loose joining relations**			**Other**		
case			list	parataxis	remnant	**Sentence head**	**Unspecified dependency**	
			dislocated		reparandum	root	dep	

Relative frequencies of the relations in the treebank

Relation	Rel. freq. (%)	Relation	Rel. freq. (%)	Relation	Rel. freq. (%)
nmod	14.6996	ccomp	1.02717	expl	0.24251
punct	13.0446	expl:pv	1.01966	goeswith	0.11675
case	12.2549	cop	0.87435	ccomp:pmod	0.0957
amod	6.56939	iobj	0.81823	remnant	0.06013
det	4.76257	nsubjpass	0.79418	advmod:tmod	0.05411
nsubj	4.63781	parataxis	0.78115	foreign	0.05111
ROOT	4.33166	auxpass	0.73556	expl:impers	0.0466
conj	4.02451	nmod:pmod	0.71501	list	0.04359
advmod	3.76847	neg	0.71	cc:preconj	0.03708
dobj	3.5941	name	0.65939	advcl:tcl	0.03658
mwe	3.04093	expl:pass	0.53814	compound	0.03658
cc	3.03893	appos	0.50106	csubjpass	0.02806
mark	2.89312	xcomp	0.46699	vocative	0.02756
acl	2.28032	nmod:tmod	0.38982	dep	0.00902
aux	2.27631	nmod:agent	0.38431	discourse	0.00802
advcl	1.48414	csubj	0.35776	reparandum	0.0005
nummod	1.34334	expl:poss	0.28811		

Example of an annotated sentence

(2) Textele acordului, anexelor, protocolului și Actului final se atașează la prezenta decizie.
(2) Texts-the agreement-of-the, annexes-of-the, protocol-of-the and Act-of-the final SE-Cl3SgAcc attach at present-the decision.
"(2) The texts of the agreement, of the annexes, of the protocol and of the Final act are attached to the present decision."



Outline

1. The treebank
    1. The corpus
    2. Its processing and annotation
    3. Syntactic annotation
        1. Principles
        2. Relations
        3. Example
2. **Data retrievable from the treebank**
    1. Statistical data with linguistic relevance
    2. Data retrievable through querying
3. Conclusions

Statistical data with linguistic relevance: focus on passive constructions (I)

	Acad	News	Biblio	EMEA	FrameNet	JRC	Lit	WIKI
auxpass	0.0081	0.0106	0.0036	**0.0151**	0.0067	0.0068	0.0038	0.0038
expl:pass	0.0036	0.0063	0.0029	0.0082	0.0007	**0.0102**	0.0026	0.0009
nsubjpass	0.0083	0.0116	0.0052	**0.0147**	0.0045	0.0110	0.0031	0.0011
csubjpass	0.0001	0.0005	0.0001	**0.0007**	0.0002	0.0002	0.0001	0.0000
nmod: agent	**0.0060**	0.0053	0.0024	0.0025	0.0027	0.0043	0.0023	0.0056

The relative frequencies of the relations connected to passive voice.

Statistical data with linguistic relevance: focus on passive constructions (II)

	Acad	News	Biblio	EMEA	FrameNet	JRC	Lit	WIKI
passive structure	0.0117	0.0170	0.0065	**0.0233**	0.0074	0.0171	0.0065	0.0048
passive subjects	0.0084	0.0121	0.0053	**0.0154**	0.0047	0.0112	0.0032	0.0011
$\frac{pass.subj}{pass.struct}$	**0.7136**	0.7121	**0.8153**	0.6609	0.6401	0.6553	0.4896	0.2239
% agent	**0.5085**	0.3131	0.3692	0.1079	0.3596	0.2499	0.3486	**1.1641**
% nsubjpass	**0.9880**	0.9574	0.9811	0.9542	0.9551	0.9814	0.9661	1

Querying the treebank

- http://bionlp-www.utu.fi/dep_search, using SETS querying system, described at http://bionlp.utu.fi/searchexpressions-new.html
- at http://lindat.mff.cuni.cz/services/pmltq/#!/home, using PML Tree Query, described at https://ufal.mff.cuni.cz/pmltq/doc/pmltq_doc.html
- at http://clarino.uib.no/iness/page?page-id=iness-main-page, with the INESS infrastructure, described at http://clarino.uib.no/iness/page?page-id=iness-documentation

Conclusions

- The existence of language resources of large size offers the researchers the opportunity to check known facts and to discover emerging tendencies.
- Besides merely reflecting various phenomena, corpora in general and treebanks in particular also inform about their frequency, which can mark either an increasing tendency or, on the contrary, rare(r) phenomena.

Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS–UEFISCDI, project number PN-II-RU-TE-2014-4-1362.