Linguistic Data Retrievable from a Treebank

Verginica Barbu Mititelu
Research Institute for Artificial
Intelligence “Mihai Drăgănescu”
Romanian Academy
vergi@racai.ro

Elena Irimia
Research Institute for Artificial
Intelligence “Mihai Drăgănescu”
Romanian Academy
elena@racai.ro

Abstract

This paper describes the Romanian treebank annotated according to the Universal Dependency principles. We present the types of texts included in the treebank, their processing phases and the tools used for doing it, as well as the levels of annotation, with a focus on the syntactic level. We briefly present the syntactic formalism used, the principles followed and the set of relations.

The perspective we adopted is the linguist’s who searches the treebank for information with relevance for the study of Romanian. (S)He can interpret the statistics based on the corpus and can also query the treebank for finding examples to support a theory, for testing hypothesis or for discovering new tendencies. We use here the passive constructions in Romanian as a case study for showing how statistical data help understanding this linguistic phenomenon. We also discuss the kinds of linguistic information retrievable and non-retrievable form the treebank, based on the annotation principles.

1. Introduction

Language resources are created both for the use of machines and for that of humans. Among the latter, several types of users can be recognised: linguistics or/and computer science researchers, teachers (of a native or foreign language), students (studying their native language or learning or studying a foreign one), or any speaker interested in various aspects of the language behaviour.

In this paper we focus on one language resource (a treebank) and show what kinds of linguistic information can be found. The language under focus here is Romanian, but the main lines of the presentation hold for any language having a treebank annotated in the same style.

In Section 2 we present the treebank: the types of texts to which the sentences in the treebank belong, processing steps, the levels of annotation, with a focus on the syntactic one: we briefly present the formalism used, the annotation principles, the inventory of relations used with emphasis on the language specific ones and exemplify with a sentence from the treebank. These data are meant as a background for understanding the rest of the paper. In Section 3 we show what kind of linguistic information can be found in the treebank, looking at passive constructions as a case study, whereas the information that cannot be found and the motivation for this are presented in Section 4. After that, we conclude the paper.

2. The Treebank

The resource which makes the topic of our paper is the Romanian treebank annotated according to the Universal Dependency (UD) guidelines[1]. A treebank is a collection of sentences annotated at the

[1] universaldependencies.org

syntactic level, i.e. syntactic relations among tokens in the sentence are marked and labelled according to their types.

2.1. The Corpus

The treebank, called RoRefTrees, contains 9522 sentences with an average length of 23 tokens. The sentences were selected from several text types: Romanian Wikipedia articles (Wiki), academic writing (Acad), newspaper articles (News), excerpts from different texts that are part of the bibliography of the Romanian Dictionary (Biblio), EMEA (Tiedemann, 2009) in Romanian, FrameNet (Baker et al., 1998) sentences translated into Romanian, the Romanian JRC-Acquis (JRC) (Steinberger, 2006), literature (Lit), medical texts (Medical). The distribution of sentences across text types is not equal, as seen in Table 1, where the Misc(ellanea) column represents a set of sentences from all the other text types (this set was firstly developed as the core of the treebank). Most sentences come from literary and legal texts. The least sentences are from medical texts, which were not among the texts we targeted at the beginning of our work, but added later on.

The tokens in the table below include both words and punctuation. The latter represents approximately 13% from the number of tokens (see Table 2). The longest sentences are in JRC and the shortest in the Biblio subcorpus (we ignored here the Misc subcorpus, given its mixed nature).

	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**

Table 1: Distribution of text types in RoRefTrees.

2.2. Texts Processing and Annotation

The texts in the treebank are tokenised, lemmatised and annotated at the morphologic and syntactic levels. Tokenisation, lemmatisation and morphologic analysis were made with the TTL tool (Ion, 2007). Although TTL uses, for tokenization, a lexicon containing "words with space", we eliminated them in a post-processing phase to comply with the UD requirements: e.g., the compound preposition "de_la" (from) is split into "de" and "la". Words with hyphens, resulted from contractions, are treated by TTL as different tokens: e.g. *n-am spus* (not-have_I said "I haven't said") is tokenised as *n-*, *am* and *spus* (the hyphen marks the elision of the vowel in the adverb of negation *nu* ("not")).

2.3. The Syntax in the Treebank

The annotation level specific to treebanks is the syntactic one. For RoRefTrees, the syntactic formalism we adopted is dependency grammar: each sentence is analysed as a tree (i.e., a directed acyclic graph). Its nodes are the words and punctuation in the sentence, while 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 that has no head in the tree is the root. Any head can have one or more dependents, or even none in the case of tree leaves.

Among the dependency grammars, we chose to work within the UD project, which aims at designing cross-linguistically consistently annotated treebanks for as many languages as possible, with the further aim of developing a parser that could run on sentences in any language.

The syntactic analysis of the sentences was made in an iterative bootstrapping way, starting from two previously available treebanks (Perez, 2014; Irimia and Barbu Mititelu, 2015), which were originally annotated following slightly different principles and sets of relations. The detailed comparison between them can be found in (Barbu Mititelu et al., 2016).

A first set of sentences (about 500) from these treebanks was manually annotated according to the principles and with the set of relations described below and, thus, a small parallel treebank was created. A correspondence table for the annotations in these parallel treebanks was created and from it a set of structural transformations in the trees were automatically learned and applied, while the conversion of relations was made by a function. The results of the automatic mapping were manually and independently checked by three linguists and, after making the necessary corrections, the sentences were used to enlarge the parallel treebank and the mapping algorithm was retrained and afterwards applied to a new set of sentences. This procedure continued until all sentences from the two treebanks were mapped to the new annotation (see Barbu Mititelu et al., 2016 for the detailed description of this process).

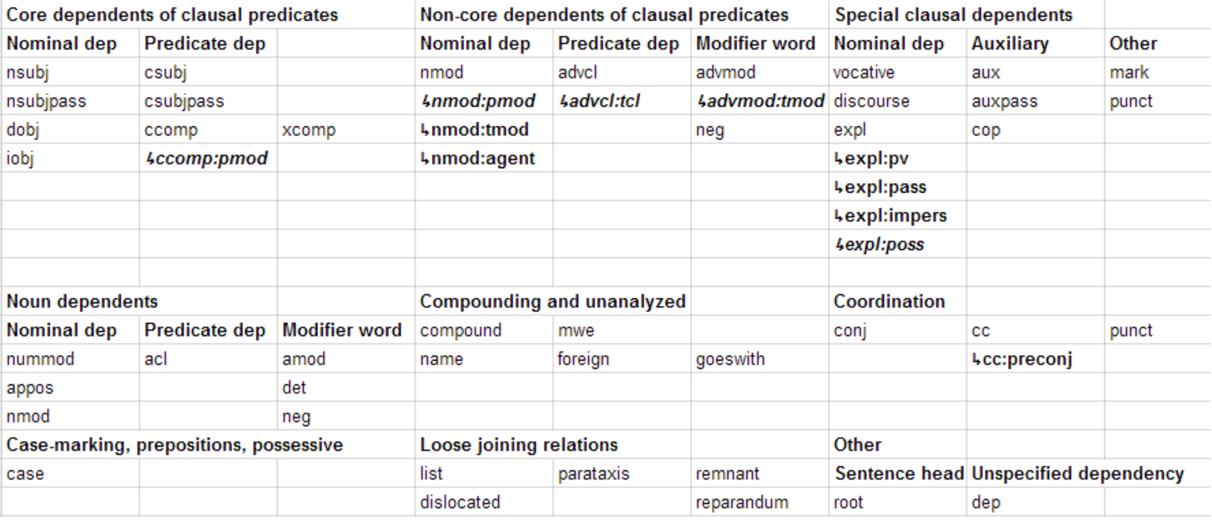
2.4. Annotation Principles

The UD annotation principles are presented on the project website and we mention them here briefly. One central principle is the treatment of function words as dependents, not as heads (except for several clear cases). A flat structure (with the first occurring element as the head and all the others as its dependents) is preferred 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.

2.5. The Set of Relations

The set of relations we used is the one in UD, which we augmented with a few language specific ones, motivated by linguistics phenomena in Romanian (see Barbu Mititelu et al., 2015 for motivations).

In UD there is a universal set of relations meant to be used for all languages. Language-specific relations are used for one or several languages displaying a certain phenomenon and are always subtypes of the universal set. In Figure 1 we put in normal font the universal relations. Their subtypes are marked by the presence of the arrow (↳). The language-specific relations used for several other languages in UD are **boldfaced**. They are used to mark the agent in passive constructions (nmod:agent), inherently reflexive verbs with a clitic pronoun (expl:pv), the reflexive clitic with a passive meaning (expl:pass), the clitic with impersonal value (expl:impers), the preconjunction (cc:preconj), and the noun with temporal value (nmod:tmod). The ***boldfaced and italic*** ones are (at least so far within UD) Romanian-specific: the obligatory prepositional object of a predicate (nmod:pmod), its clausal equivalent (ccomp:pmod), time adverbials (advcl:tcl), time adverbs (advmod:tmod), possessive dative (expl:poss).



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	

Figure 1: Syntactic relations used in RoRefTrees.

The relative frequency of all these relations in RoRefTrees is presented in Table 2. The most frequent relation is nmod (marking the nominal modifier of a word). Punctuation comes next and prepositions (marked with the case relation) after it. Further discussions about the interpretation of data in this table can be found in section 3.1.

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		

Table 2: The relative frequencies of the relations in RoRefTrees.

2.6. Example

A tree from RoRefTrees is presented in Figure 2. It renders the syntactic analysis of the sentence:

(1) (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."

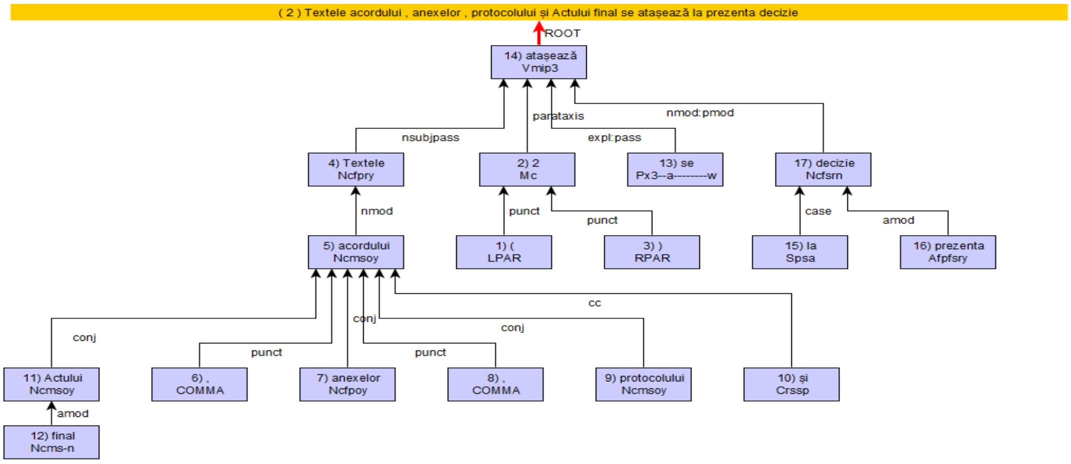


Figure 2: A tree from RoRefTrees.

This is a sentence with a verbal root (*atașează*), a reflexive clitic with passive value (*se*), a nominal subject of a passive verb marked as such (*Textele*). The subject has four coordinated nominal modifiers (*acordului*, *anexelor*, *protocolului*, *Actului*), out of which only the first is analysed as a dependent of the subject, while the others are analysed as conjuncts of it. The commas between the coordinated elements are also attached to the first conjunct, just like the coordinating conjunction (*și*). The preposition (*la*) is attached via the `case` relation to the noun it precedes. The number of the law article (2) (this sentence is from the JRC subcorpus) is attached as `parataxis` to the root of the tree. All punctuation is attached to the head via the relation `punct`: final punctuation to the root of the tree, the parentheses to the element they isolate from the rest of the sentence.

3. What data can a linguist find in the treebank?

A treebank can offer precious information to a linguist in two ways: statistically and by searching it. We will consider them in turn below.

3.1. Let the numbers talk!

In this section we focus on the linguistic relevance of the figures and per cents in the tables above and below. The relative frequencies of specific phenomena either with respect to the whole treebank or to subparts of it can offer information that is difficult to have access to without a treebank. They offer the linguists solid ground for quantitative statements that were difficult to make before the existence of corpora. We will use the passive construction in Romanian as a case study and in the rest of this subsection we will analyse the data pertinent to it found in RoRefTrees.

The passive voice has two possible realisations in Romanian:

- with auxiliary:

(2) Copilul este sărutat de părinte.
Child-the is kissed by parent.
"The child is kissed by his parent."

The passive auxiliary is *este* in this example, the third person singular of the verb *a fi* ("to be").

- with reflexive clitic:

(3) Contractul se va semna mâine de către reprezentanții celor două instituții.
Contract-the SE-Cl3SgAcc will sign tomorrow of towards representatives-the those-of two institutions.
"The contract will be signed tomorrow by the representatives of the two institutions."

The reflexive clitic with passive meaning is *se* in such constructions.

The relations identifying the passive in our treebank are: `auxpass` (the passive auxiliary), `refl:pass` (the reflexive clitic with a passive value), `nsubjpass` (the nominal subject of a passive verb), `csubjpass` (the clausal subject of a passive verb), `nmod:agent` (the nominal agent complement of the verb). A clausal realization of the agent complement is possible in Romanian, but it never occurred in our treebank. The first two relations are mandatory for a sentence to be interpreted as passive (but they cannot co-occur). The others are optional: the absence of the subject (either nominal or clausal) is possible given the fact that Romanian is a pro-drop language, whereas the absence of the agent nominal is "a vivid iconic manifestation of the most salient functional-pragmatic feature of the passive voice – agent suppression" (Givón, 2001: 126).

The data in Tables 2 and 3 shows several things related to passive. First, no relation specific to passive is too frequent in any text type. From this we can conclude that active voice is much more frequent than passive voice. Second, the distribution of the passive construction across the text types shows that the passive is the most frequent in the EMEA texts. According to linguistic literature on this topic (see Quirk et al., 1984: 166; Givón, 2001: 125; among others), informative texts favour passive constructions; the data in Table 4 shows the same tendency: EMEA sentences, i.e. scientific texts, have the highest relative frequency of passive structures, while Lit and Biblio, i.e. imaginative texts, have the lowest relative frequency of passive constructions. What is intriguing is that Wikipedia texts, which belong to the category informative rather than imaginative, show a lower frequency of passives than imaginative texts. A motivation for this will probably be found when a semantic analysis of the sentences from Wikipedia is made.

Third, the passive voice with auxiliary is much more frequent than the reflexive passive: the relative frequency of `auxpass` is higher than the relative frequency of `expl:pass`. In Table 3 we can see that this statement holds true for all text types in the treebank, with the only exception of the JRC sentences, in which the impersonal reflexive form prevails.

Fourth, the passive subjects are also most frequently realised in the EMEA sentences (0.0154) (see the line "passive subjects" in Table 4). However, the lexicalization of the subject in passive sentences happens most frequently in the Biblio subcorpus. The explanation resides in the fact that these sentences were selected by the dictionary editors to serve as examples of the usage of a lexical unit, so they must be characterized by semantic and syntactic completeness, coherence, cohesion, lack of ambiguity.

Fifth, the most frequent type of subject realisation is the nominal one (in more than 95% of the cases: see line "%nsubjpass" in Table 4) and its relative frequency is the highest in the Acad subcorpus. This correlates with the data in Tabel 2, which show higher relative frequencies for `nsubj` and `nsubjpass` than for `csubj` and `csubjpass`.

Sixth, the relative frequency of the realisation of agents in passive structures is below 50%, with the highest in Acad: 0.5085. However, one can see that in Wikipedia texts the relative frequency of the realisation of agent is 1.1641. This is informative of the fact that nominal agents occur in constructions that are not syntactically passive, but carry a passive meaning: for instance, the verbal nominalisation in this example:

(4) Sărutarea copilului de către părinte ....
Kissing-the child-the-of of towards parent ....
"The kissing of the child by his parent …."

The noun *sărutarea* ("the kissing") preserves the semantic arguments of the verb it is derived from: the agent and the patient. The former is realised in the same morpho-syntactic form as in the passive voice, namely with the compound preposition "de către" (*by*).

	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

Table 3: The relative frequency of relations connected to passive voice in RoRefTrees subcorpora.

	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
$\frac{\boldsymbol{passive\ subjects}}{\boldsymbol{passive\ structures}}$	**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

Table 4: Further relative frequencies connected to passive voice in RoRefTrees.

3.2. What types of searches can be made in the treebank?

Besides analysing the figures in the statistics drawn from the treebank, the linguist can also search for various structures and their instantiation in it. RoRefTrees are available for download on the UD website, with the content from the last release. The treebank can also be queried online using different tools: at 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 (Rosén et al., 2012) infrastructure, described at http://clarino.uib.no/iness/page?page-id=iness-documentation.

One can search a treebank for a multitude of linguistically relevant data. Their analysis reflects the grammatical theory that was used for annotation. We present below several examples of searches:

- the arguments of a certain verb: one can extract all core dependents of the respective verb, even with the aim of creating a valence dictionary of the verbs in the treebank; these core dependents are words linked to the respective verb by any of the relations `nsubj`, `nsubjpass`, `csubj`, `csubjpass`, `dobj`, `iobj`, `ccomp`, `ccomp:pmod`; besides them, one must also consider `nmod:pmod` and `nmod:agent` relations, although they are classified under non-core dependents in Figure 1;

- the parts of speech a certain syntactic function can be realised by: for example, what parts of speech the root of a clause can be; in RoRefTrees one will find verbs, interjections, nouns, adjectives and adverbs as roots. If Romanian traditional grammar has the notions of predicative interjections and adverbs, so these two parts of speech are no surprise among the results, then the adjective and nouns are unexpected roots in non-elliptical structures, but this is the result of the convention used for annotating the copula verb *a fi* ("to be"): a dependent on the adjective or noun, linked by the `cop` relation: in Figure 3 we present the analysis of the adjective *frumoasă* from the sentence in (5) as the root of the sentence.

(5) Fata este frumoasă.
Girl-the is beautiful.
"The girl is beautiful."

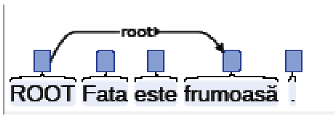


Figure 3. A sentence with an adjectival root.

- the words realising a syntactic function for a certain word: one may want to identify the semantic restrictions on a certain argument of a verb; this can be done by analysing all the words filling that position in the argument structure of the respective verb in the treebank;

- the parts of speech between which a certain syntactic relation establishes: for example, `iobj`, which is found in our treebank as occurring between nouns, pronouns as dependents and verbs, adjectives or interjections as heads. The analysis can go even further: one can look at various morphologic characteristics of these parts of speech, such as case for nouns or pronouns;

- the word order (even in different types of sentences, such as declarative, interrogative, exclamatory, affirmative, negative); an interesting study for a language with relative free word order would be the position of the subject, when lexicalised: pre- or post-verbal position.

- etc.

4. What Cannot Be Found in RoRefTrees?

The conventions in the formalism adopted for creating the treebank have consequences in the type of information retrievable from the treebank. We discuss several disadvantages of the annotation here.

When designing the set of relations to be used in the syntactic annotation (within UD), both structure and function were considered. Some relations clearly reflect the way dependents function in the sentence: `dobj`, `iobj`, etc. Others reflect rather the morphologic components: see `nmod` and `advmod` relations: the former functionally corresponds to an adverbial when it attaches to a verb, adjective or an adverb, but when attaching to a noun, it corresponds to an attribute; the latter is an adverb or adverbial phrase that serves to modify the meaning of its head. There are others that combine both aspects: `nsubj`, `csubj`: they are used for the same syntactic position (a subject), but the former is used for nominals filling this position, while the latter for clauses.

Sometimes, the same relation is used to link both arguments and adjuncts to their heads: e.g. `advmod`. It is impossible to automatically distinguish between adverbs that are arguments, as in (6), and those that are adjuncts, as in (7), as the same relation (`advmod`) links them to their head.

(6) El se poartă frumos.
He Se-Cl3SgAcc behaves beautifully.
“He behaves himself.”
(7) El cântă frumos.
He sings beautifully.
“He sings beautifully.”

In Figure 1, one can notice that the clausal realisation of both the direct and indirect objects is linked to the head by the same relation, `ccomp`, which means that no distinction between the two positions can be made automatically. One way of disambiguating this relation is to look for a `dobj` or `iobj` of the same head: as there cannot be two `dobj` or `iobj` relations of the same head, the co-occurrence between a `dobj` and a `ccomp`, for instance, will help infer the fact that the subordinate clause fills the indirect object slot of the head argument structure. Otherwise, we cannot see another way for telling the values of the `ccomp` apart.

5. Conclusions

Nowadays, when language resources are being created and their size is in continuous increase, the researchers interested in the study of a language focus more on these resources, search them for known facts and new 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 phenomena.

We presented above the Romanian treebank annotated according to UD conventions and discussed about several information types a linguist can search for and find in it. Others remain covert and other solutions need to be found for spotting them in the treebank.

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.

References:

Baker, C. F., Fillmore, Ch. J., Lowe, J. B. (1998). The Berkley FramNet Project. *Proceedings of the 17th International Conference on Computational Linguistics* - Volume 1.

Barbu Mititelu, V., Irimia, E., Mărănduc, C. (2015). Universal and Language-specific Dependency Relations for Analysing Romanian. *Proceedings of the Third International Conference on Dependency Linguistics (DepLing2015)*, August 24-26, Uppsala, Sweden.

Barbu Mititelu, V., Ion, R., Simionescu, R., Irimia, E., Perez, C.-A. (2016). The Romanian Treebank Annotated According to Universal Dependencies. *Proceedings of HrTAL*, September 29 – October 1, Dubrovnik, Croatia.

Givón, T. (2001). *Syntax: An Introduction*. Vol. II. Amsterdam/Philadelphia: John Benjamins.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD thesis, Romanian Academy (in Romanian).

Irimia, E., Barbu Mititelu, V. (2015). Building a Romanian Dependency Treebank. *Corpus Linguistics 2015*, Lancaster, UK, 21-24 July 2015.

Nivre, J., Hall, J. Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, May 24 – 26, 2006, Genoa, Italy.

Perez, C.-A. (2014). *Resurse lingvistice pentru prelucrarea limbajului natural*, PhD thesis, A.I. Cuza University of Iasi.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Rosén, V., De Smedt, K., Meurer, P., Dyvik, H. (2012). An Open Infrastructure for Advanced Treebanking. In: Hajič, J., De Smedt, K., Tadić, M., Branco, A. (eds.) *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, May 21-27, Istanbul, Turkey.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiș, D., Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. May 24-26, Genoa, Italy.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) *Recent Advances in Natural Language Processing* (vol V). Amsterdam/Philadelphia: John Benjamins.