Adding semantics to UD markers

Tudor Voicu
Tudor Vianu National
High School of Computer Science
Bucharest, Romania
tudor.c.voicu@gmail.com

Verginica Barbu Mititelu
Romanian Academy
Research Institute for Artificial Intelligence
Bucharest, Romania
vergi@racai.ro

Relevant UniDive working groups: WG1

1 Introduction

The initial steps towards the enrichment of the Romanian Reference Treebank with semantic information are presented below. The focus is on conjunctive locutions: they have a double role: they contribute to text cohesion and are lexical devices for expressing the logical relations between units of meaning in a text. We annotate their occurrences in the Romanian Reference Treebank (RRT) (Barbu Mititelu, 2018) with the discourse relations defined in the manual for Penn Discourse Treebank version 3.0. The results of the double annotation are discussed, the types of relations associated with each locution are presented and the way in which this level of annotation is added to the existing resource is explained.

2 State of the art

RRT is a collection of 9,523 sentences from several genres: legal, news, fiction, medical, non-fiction, academic, FrameNet translations and Wikipedia, with an unbalanced distribution. The corpus is morpho-syntactically annotated according to the principles of Universal Dependencies[1] (UD) (de Marneffe et al., 2021), being available in CoNLL-U format[2].

UD principles of syntactic annotation give primacy to content words, i.e., dependency relations hold directly between content words, not mediated by function words, which are (always) annotated as leaves in the trees. Content words are nouns, adjectives, verbs (but not auxiliaries) and adverbs, while the other parts of speech are function words, conjunctions included. They are words that help clarify the syntactic and/or semantic relation between other words or phrases in the sentence. Coordinating conjunctions are annotated with the relation `cc` holding between them and the last conjunct in the coordination (see Figure 1 representing the analysis of the sentence *Dați-mi voie **și** vă spun, cuvânt cu cuvânt.* “Allow me and I will tell you, word by word.”). Subordinating conjunctions are annotated with the relation `mark` holding between them and the head of the subordinate clause it introduces (see Figure 2 representing the analysis of the sentence *Unul dintre ecrane s-a întrerupt **deoarece** el a lovit un avocat.* “One of the screens turned off because he hit a lawyer.”).

Conjunctive locutions are treated as multiword tokens, i.e., they are segmented into components and only the first component establishes the relation `mark` with the verbal head, while the other component(s) is/are dependent(s) of the first one and is/are linked by means of the `fixed` relation: see Figure 3.

In traditional Romanian linguistics, there are some conjunctive locutions that end with the conjunction *să*, which is the marker of the conjunctive mood in Romanian: e.g., *fără să* ‘without SĂ’, *în loc să* ‘instead of SĂ’, etc. For a consistent annotation of verbs in the conjunctive mood in RRT, *SĂ* is always attached by the `mark` relation to the verb, including situations when it is also part of such conjunctive locutions. This consistency, however, leads to an inconsistent treatment of traditionally coined conjunctive locutions: those whose last component is not *SĂ* are annotated as shown in Figure 3 (representing the analysis of the sentence *Echipajul său de opt oameni a pierit de asemenea **în timp ce** se zbătea să salveze echipajul Santampa.* “is crew of eight people also died while striving to save the Santampa crew.”), while those ending in *SĂ* are actually split, with this last component not annotated as part of the locution (see Figure 4 representing the analysis of the string ***în loc să*** *mă plimb* “instead I walk”). Our annotation (see Figure 5) is able to render the whole locutions, as attested in grammars and dictionaries of the language.

3 Inventory of semantic relations

For the semantic annotation of conjunctions and conjunctive locutions in RRT we chose the set of

[1] https://universaldependencies.org/

[2] https://universaldependencies.org/format.html

discourse relations defined for the annotation of the Penn Discourse Treebank[3] (PDTB) (Prasad et al., 2019). These relations are organized hierarchically, on three levels, on the most general one being the relations Temporal, Contingency, Comparison, and Expansion. They are further refined in the next two levels, also considering implicit beliefs (epistemic knowledge) and speech acts associated with the arguments of the respective relations.

4 Annotation results

Each occurrence of the conjunctive locutions is annotated by two annotators. Each relation is established between two arguments. An argument is usually a clause, in syntactic terms, i.e., a syntactic unit organized around a verb. For each occurrence of a conjunction, the two arguments that it links are identified and then a relation from the PDTB hierarchy is assigned to it: only relations of the third level were assigned, i.e. the most specific ones; only when there is no relation on this level, a second level one was chosen. In case the context is insufficient, the annotation is skipped.

We show the relations expressed by the so far annotated conjunctions in Table 1. The values presented here are those established by the two annotators after discussing together the cases when they annotated differently. We can see that locutions can have more than one meaning. Out of the four locutions already annotated, only one (*înainte de* 'before') is monosemous (it is annotated only with the relation TEMPORAL:ASYNCHRONOUS:PRECEDENCE), while the others have more senses; nevertheless, there is one value that prevails for each of them: TEMPORAL:SYNCHRONOUS for *în timp ce* 'while', CONTINGENCY:CAUSE:REASON for *pentru că* 'because' and TEMPORAL:ASYNCHRONOUS:SUCCESSION for *după ce* 'after'. There is also another relation that stands out among the others for each conjunction: COMPARISON:CONTRAST for *în timp ce* and CONTINGENCY:CAUSE+BELIEF:REASON+BELIEF for *pentru că*. The other relations annotated only occur in a few sentences. This polysemous nature of the conjunctions is one of the sources of disagreements between annotators.

A special situation is represented by the locution *după ce*, which sometimes seems to express two relations simultaneously: TEMPORAL:ASYNCHRONOUS:SUCCESSION and CONTINGENCY:CAUSE:REASON.

Although the discourse relations expressed by these four locutions are quite different, we, of course, expect cases when different conjunctions will express the same relation, thus showing synonymy.

The annotated discourse relations are added on the tenth column of the CoNLL-U file: the first element of the locution gets a numerical identifier followed by a colon and the name of the relation; the other elements only get the same number as the first element and the name of the relation remains unspecified. Additionally, when more conjunctive locutions occur within a sentence, each of them is assigned another number. Figure 5 shows the CoNLL-U format of a sentence. The annotated locutions are highlighted in green. In their last column, we notice the number of the locution and, on its first component, the name of the relation is also added. The same number assigned to different tokens means that they are components of the same locution. This way of annotation is similar to the one used in the PARSEME corpora (Savary et al., 2017).

5 Conclusions

We have presented here the framework and the first steps taken towards the enrichment of RRT with discourse relations from the PDTB 3.0 hierarchy. The work is only at the beginning, thus the reported results are scarce.

The data can give valuable insights into the behaviour of Romanian conjunctions with respect to the discourse relations they express, their default readings and also possible misinterpretations (in cases revealed by the double annotation we are carrying out). Comparisons between conjunctives considered equivalent in different languages (i.e., those for which such annotations do exist) will thus be possible and can help the translation process and text interpretation.

References

Verginica Barbu Mititelu. 2018. Modern syntactic analysis of Romanian. In *Clasic şi modern în cercetarea filologică românească actuală*, pages 67–78, Iași, Romania.

[3] https://catalog.ldc.upenn.edu/LDC2019T05

Conjunction	**Frequency**	**Annotation**
în timp ce	29	TEMPORAL:SYNCHRONOUS
‘while’	18	COMPARISON:CONTRAST
	2	COMPARISON:CONCESSION:ARG2-AS-DENIER
	1	COMPARISON:SIMILARITY
	1	insufficient context
pentru că	35	CONTINGENCY:CAUSE:REASON
‘because’	10	CONTINGENCY:CAUSE+BELIEF:REASON+BELIEF
	4	CONTINGENCY:CAUSE+SPEECHACT:REASON+SPEECHACT
	3	CONTINGENCY:CAUSE+SPEECHACT:RESULT+SPEECHACT
	2	CONTINGENCY:CAUSE:RESULT
înainte de	35	TEMPORAL:ASYNCHRONOUS:PRECEDENCE
‘before’		
după ce	48	TEMPORAL:ASYNCHRONOUS:SUCCESSION
‘after’	38	TEMPORAL:ASYNCHRONOUS:SUCCESSION| CONTINGENCY:CAUSE:REASON
	1	TEMPORAL:ASYNCHRONOUS:PRECEDENCE| TEMPORAL:ASYNCHRONOUS:SUCCESSION

Table 1: The discourse relations expressed by the annotated conjunctions.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. *Penn Discourse Treebank Version 3.0*. Linguistic Data Consortium.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

A Appendix

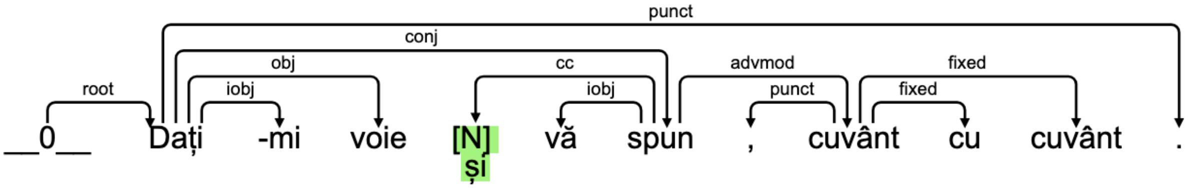


Figure 1: Representation of dependency relations: function words are always dependents: the case of coordinating conjunctions.

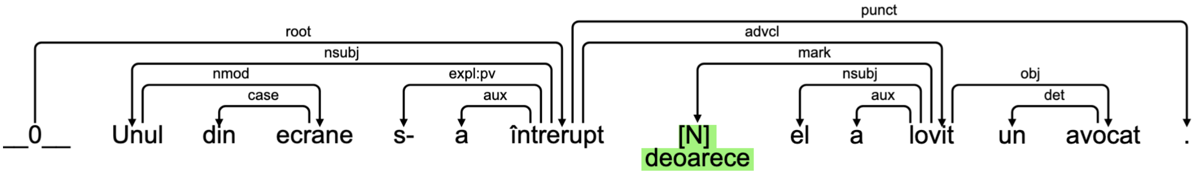


Figure 2: Representation of dependency relations: function words are always dependents: the case of subordinating conjunctions.

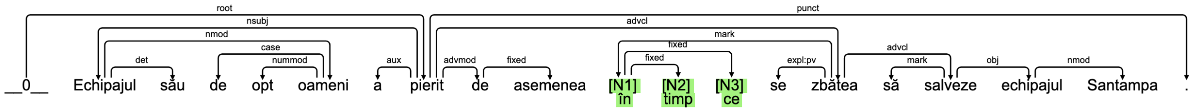


Figure 3: Representation of dependency relations: annotation of conjunctive locutions.

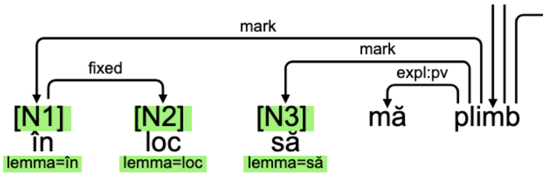
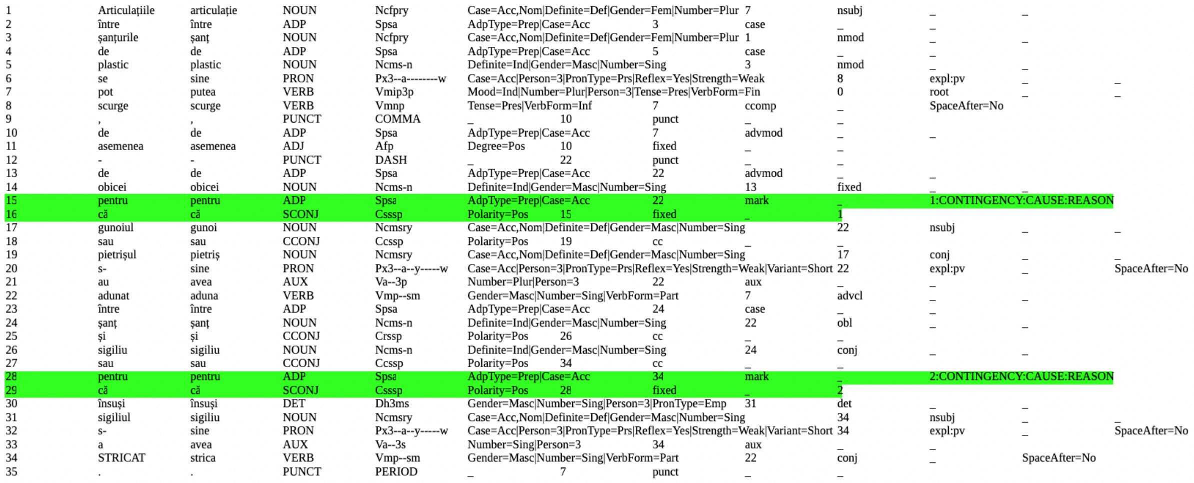


Figure 4: Representation of dependency relations: annotation of conjunctive locutions.



```
1   Articulațiile  articulație  NOUN   Ncfpry           Case=Acc,Nom|Definite=Def|Gender=Fem|Number=Plur  7  nsubj  _  _
2   între          între        ADP    Spsa             AdpType=Prep|Case=Acc  3  case  _  _
3   șanțurile      șanț         NOUN   Ncfpry           Case=Acc,Nom|Definite=Def|Gender=Fem|Number=Plur  1  nmod  _  _
4   de             de           ADP    Spsa             AdpType=Prep|Case=Acc  5  case  _  _
5   plastic        plastic      NOUN   Ncms-n           Definite=Ind|Gender=Masc|Number=Sing  3  nmod  _  _
6   se             sine         PRON   Px3--a--------w  Case=Acc|Person=3|PronType=Prs|Reflex=Yes|Strength=Weak  8  expl:pv  _  _
7   pot            putea        VERB   Vmip3p           Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin  0  root  _  _
8   scurge         scurge       VERB   Vmnp             Tense=Pres|VerbForm=Inf  7  ccomp  _  SpaceAfter=No
9   ,              ,            PUNCT  COMMA            _  10  punct  _  _
10  de             de           ADP    Spsa             AdpType=Prep|Case=Acc  7  advmod  _  _
11  asemenea       asemenea     ADJ    Afp              Degree=Pos  10  fixed  _  _
12  -              -            PUNCT  DASH             _  22  punct  _  _
13  de             de           ADP    Spsa             AdpType=Prep|Case=Acc  22  advmod  _  _
14  obicei         obicei       NOUN   Ncms-n           Definite=Ind|Gender=Masc|Number=Sing  13  fixed  _  _
15  pentru         pentru       ADP    Spsa             AdpType=Prep|Case=Acc  22  mark  _  1:CONTINGENCY:CAUSE:REASON
16  că             că           SCONJ  Csssp            Polarity=Pos  15  fixed  _  1
17  gunoiul        gunoi        NOUN   Ncmsry           Case=Acc,Nom|Definite=Def|Gender=Masc|Number=Sing  22  nsubj  _  _
18  sau            sau          CCONJ  Ccssp            Polarity=Pos  19  cc  _  _
19  pietrișul      pietriș      NOUN   Ncmsry           Case=Acc,Nom|Definite=Def|Gender=Masc|Number=Sing  17  conj  _  _
20  s-             sine         PRON   Px3--a--y-----w  Case=Acc|Person=3|PronType=Prs|Reflex=Yes|Strength=Weak|Variant=Short  22  expl:pv  _  SpaceAfter=No
21  au             avea         AUX    Va--3p           Number=Plur|Person=3  22  aux  _  _
22  adunat         aduna        VERB   Vmp--sm          Gender=Masc|Number=Sing|VerbForm=Part  7  advcl  _  _
23  între          între        ADP    Spsa             AdpType=Prep|Case=Acc  24  case  _  _
24  șanț           șanț         NOUN   Ncms-n           Definite=Ind|Gender=Masc|Number=Sing  22  obl  _  _
25  și             și           CCONJ  Crssp            Polarity=Pos  26  cc  _  _
26  sigiliu        sigiliu      NOUN   Ncms-n           Definite=Ind|Gender=Masc|Number=Sing  24  conj  _  _
27  sau            sau          CCONJ  Ccssp            Polarity=Pos  34  cc  _  _
28  pentru         pentru       ADP    Spsa             AdpType=Prep|Case=Acc  34  mark  _  2:CONTINGENCY:CAUSE:REASON
29  că             că           SCONJ  Csssp            Polarity=Pos  28  fixed  _  2
30  însuși         însuși       DET    Dh3ms            Gender=Masc|Number=Sing|Person=3|PronType=Emp  31  det  _  _
31  sigiliul       sigiliu      NOUN   Ncmsry           Case=Acc,Nom|Definite=Def|Gender=Masc|Number=Sing  34  nsubj  _  _
32  s-             sine         PRON   Px3--a--y-----w  Case=Acc|Person=3|PronType=Prs|Reflex=Yes|Strength=Weak|Variant=Short  34  expl:pv  _  SpaceAfter=No
33  a              avea         AUX    Va--3s           Number=Sing|Person=3  34  aux  _  _
34  STRICAT        strica       VERB   Vmp--sm          Gender=Masc|Number=Sing|VerbForm=Part  22  conj  _  SpaceAfter=No
35  .              .            PUNCT  PERIOD           _  7  punct  _  _
```

Figure 5: Representation of discourse relations in the CoNLL-U file.