Romanian Inherently Adpositional Verbs (IAVs)

Mihaela Cristescu
University of Bucharest
Faculty of Letters

Verginica Barbu Mititelu
Romanian Academy
Research Institute for Artificial Intelligence

Relevant UniDive working group: WG1

1 Introduction

We aim at defining and refining the class of Inherently Adpositional Verbs (IAVs) in the Romanian language, as well as in others, by offering a solid classification and reliable diagnostic tests, that can be further applied to other languages. In the process of enriching the latest version of our PARSEME corpus with this experimental and optional category, we noticed some challenges in annotating it. At the same time, such expressions are widely spread in our corpus: they represent a third of the total number of VMWEs occurring therein. This makes them an important phenomenon to be made explicit in a corpus.

2 State of the art

PARSEME guidelines describe the class of IAVs as consisting of "a verb or a VMWE and an idiomatic selected preposition or a postposition that is always required or, if absent, changes the meaning of the verb or of VMWE significantly"[1]. They correspond to what is called in English *prepositional verb*s[2].

During the annotation campaigns for the PARSEME shared tasks, this was a special experimental and optional category. Thus, to the best of our knowledge, IAVs have been annotated so far only for the English corpus in the PARSEME Shared Task 1.1 and for the Italian and Irish corpora in the PARSEME Shared Task 1.2. The IAVs annotation process for the English corpus, as well as the challenges raised by the confusion between the IAV and the verb-particle construction (VPC) classes are discussed by Walsh et al. (2018). As for Italian, the IAVs have been annotated "non-systematically on an experimental basis" (Ramisch et al., 2020). This class has been compared to light verb constructions (LVCs), as they are considered borderline cases in Italian (Monti and di Buono, 2019). Similarly, the annotation of IAVs raised some challenges for Irish, as many LVCs select for a specific preposition which is often included as an integral part of the construction (Walsh et al., 2020). Furthermore, "frequent contractions of prepositions with personal pronouns make it hard to annotate IAVs in the Irish corpus" (Ramisch et al., 2020).

3 PARSEME-Ro annotated with IAVs

The PARSEME-Ro corpus was annotated with VMWEs in two phases. In a first step (consisting of all three annotation phases pertaining to the participation in the three editions of the shared tasks), universal and quasi-universal types of VMWEs were manually identified and classified (LVCs.full, LVCs.cause, VIDs and IRVs). In a second, recent step, IAVs have been annotated in PARSEME-Ro. This time, we started from a manually compiled list of 1725 prepositional verbs (Geană, 2013). The annotation was automatic, followed by manual validation and correction (Barbu Mititelu et al., 2022). In (1) we show an example of a IAV annotated in the PARSEME-Ro corpus.

(1) Impactul   a   **dus**   **la**
    Impact.DEF. have.3.SG. lead.PART. to
    decesul   lui Tiberiu.
    death.DEF. of Tiberiu
    'The impact lead to Tiberiu's death.'

Table 1 shows that two thirds of the automatically annotated IAV are actually correct IAVs and that the decision to automatically annotate them was a time saving one. They represent 94.47% of the IAVs that should have been annotated, i.e. of the cases called "gold IAVs" in the table. Each individual initial manual validation covers almost 90% of all correct cases: see the last column of lines two and three in Table 1.

Information about the VMWEs types and their frequency in the PARSEME-Ro corpus is provided

---

#	Total IAVS	correctly annotated	
		#	%
AUTO annot.	4,686	3,128	66.75
annot. 1	3,462	3,085	89.11
annot. 2	3,519	3,185	90.5
both annots	-	2,981	
gold IAVs	-	**3,311**	

Table 1: General statistics of the IAV annotation process

Type	Number
IRV	3.826
LVC.cause	182
LVC.full	516
VID	1.644
IAV	3.311
TOTAL	9479

Table 2: Number of VMWEs of each type

in Table 2, which shows that a third of the VMWEs in the corpus is represented by IRVs and another third by IAVs.

4 IAVs annotation

The PARSEME guidelines for annotating IAVs propose a single test designed to distinguish this class from non-IAVs[3]: if in response to a declarative sentence containing the potential IAV, a question cannot be asked about the circumstances of the verbal event using the verb, but not the adposition, then the combination verb + adposition is annotated as IAV.

When testing the IAVs from our corpus against the PARSEME criterion, we noticed that for some of them the test does hold, as one cannot ask about the circumstances of the verbal event using the verb only, without the adposition, as shown in (2), which is a grammatically incorrect question in Romanian.

(2)　*Când **ne**　　**bazăm**?
　　　When us.ACC rely.PRES.1.PL?
　　　'When do we rely?'

However, there are other VMWEs annotated as IAVs in our corpus for which the test does not hold. Asking a question like (3) is grammatically correct.

---

[3]The test is an adaptation of STREUSLE's guideline on prepositional verbs by Nathan Schneider and Meredith Green: https://github.com/nschneid/nanni/wiki/Prepositional-Verb-Annotation-Guidelines

(3)　Când ai　　　**aplicat**?
　　　When have.2.SG. apply.PART.?
　　　'When did you apply?'

The difference between the verbs in (2) and in (3) is the following: the latter verb (*aplica* 'apply') can be used in a context in which the prepositional phrase occurs in the context preceding the question, which makes it possible not to repeat the prepositional phrase in the question; the former verb (*se baza* 'rely') cannot be used like this: the prepositional phrase must occur in the question, even if previously mentioned in the context.

Given these remarks, we propose to split our IAVs into two main categories, based on the PARSEME criterion.

5 Prepositional verbs in traditional grammar

A terminological clarification is needed here. In the literature, prepositions are considered a semi-lexical category (Corver and van Riemsdijk, 2013). For Romanian, two types of prepositions have been identified, namely the lexical prepositions and the functional ones (Pană Dindelegan, 2013). Prepositions belonging to the latter type have no semantics and are specialized for syntactic functions or relations displaying certain semantic or morphological characteristics: e.g., the functional preposition *pe* introduces the direct object having certain semantic characteristics (such as [+person], [+animate], etc.) or morphological ones (see the occurrence of certain types of pronouns in this syntactic position). There are only four such prepositions in Romanian and only a few of all their possible uses are considered functional (Pană Dindelegan, 2013, 454-7). All the others belong to the former type. In this paper we do not use the terms *functional* and *semi-lexical* with these meanings.

Both subclasses of IAVs described in Section 4 are presented as a single category in the Romanian grammars, under the umbrella of *subcategorized prepositions* (Pană Dindelegan, 2013, 457). They are considered to share characteristics both with the functional and the (semi-)lexical prepositions.

Unfortunately, the cases when a verb requires a certain preposition are not explicitly described in the lexicographic definitions. One counter-example for Romanian is the *Dictionary of Verbal Contexts*[4] (Barbu, 2017), which comprises a list

---

[4]https://dcv.lingv.ro/

of 600 verbs and over 2,000 meanings with one or more valency patterns.

6 IAVs classification

While annotating the IAVs, we noticed the need of refining the classification of IAVs in our language, as there are at least two main subclasses, based on the adposition behaviour. Our proposal is to split the class of IAVs into two main types, namely *IAV.functional* and *IAV.semi-lexical*, based on the observance of the PARSEME criterion. For the IAVs in the first category, the PARSEME test holds, as one cannot ask about the circumstances of the verbal event using only the verb, but not the preposition. This class is further divided into 3 subtypes, based on the semantics of the verb. However, the IAVs in the second class do not pass the test, as one can ask a question using only the verb, without changing the meaning of verb or altering the grammaticality of the structure. Therefore, we propose the following classification of the IAVs in our language:

1) IAV.functional – this subclass contains verbs which never occur without the required preposition. We can further identify three subtypes, based on the semantics of the verb:

1a) The verb is polysemous, but it always requires the same preposition, irrespective of its meaning: *a se baza pe* 'rely on', *a depinde de* 'depend on', *a consta în/din* 'consist in/of';

1b) The verb is polysemous, but it requires a certain preposition only when used with a certain meaning: *a da (peste)* 'run into', *a aparține (de)* 'belong to', *a conta (pe)* 'count on';

1c) The verb is symmetrical/reciprocal and when used with a prepositional object (and not with a plural noun as subject, for example), it always requires the preposition *cu* corresponding to the English *with*: *a merge (cu)* 'match with', *a coincide (cu)* 'coincide with', *a colabora (cu)* 'collaborate with';

2) IAV.semi-lexical – this type comprises verbs that normally require a specific preposition. They are polysemous and when used with certain meanings, they are always followed by a specific preposition: *a aplica (la/pentru)* 'apply for', *a adera (la)* 'adhere to', *a-i păsa (de)* 'care about'.

7 Conclusions and future work

We consider that the IAV class of VMWEs as defined in PARSEME needs a more fine-grained anal-

ysis. As expected, the subtype of *IAV.functional* displays a limited set of elements, that represent the actual IAVs, if we follow the PARSEME test. Nevertheless, annotating the *IAV.semi-lexical* category will prove very valuable for our language, at least from one perspective, i.e. the acquisition of Romanian as a foreign/second language.

Continuing the PARSEME custom of testing classes and subclasses against data in more languages before coining them officially, the next step we envisage is collaborating with teams working on IAVs for other languages, so as to share findings.

So far, only verbal IAVs have been annotated in PARSEME-Ro, while inherently adpositional VMWEs (IAVMWEs) are left for further investigations, but we expect to encounter similar problems in their case too.

References

Ana-Maria Barbu. 2017. *Dicţionar de contexte verbale [Dictionary of verbal contexts]*. Editura Universităţii din Bucureşti.

Verginica Barbu Mititelu, Mihaela Cristescu, Maria Mitrofan, Bianca-Mădălina Zgreabăn, and Elena-Andreea Bărbulescu. 2022. A Romanian treebank annotated with verbal multiword expressions. In *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, pages 137–145, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.

Norbert Corver and Henk van Riemsdijk. 2013. *Semi-lexical Categories: The Function of Content Words and the Content of Function Words*. De Gruyter Mouton.

Ionuţ Geană. 2013. *Construcţii verbale prepoziţionale în limba română*. Editura Universităţii din Bucureşti.

Johanna Monti and Maria Pia di Buono. 2019. Parseme-it: an italian corpus annotated with verbal multiword expressions. *Italian Journal of Computational Linguistics*, 5–2:61–93.

Gabriela Pană Dindelegan. 2013. *The Grammar of Romanian*. Oxford University Press.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the*

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 107–118, online. Association for Computational Linguistics.

Abigail Walsh, Claire Bonial, Kristina Geeraert, John P. McCrae, Nathan Schneider, and Clarissa Somers. 2018. Constructing an Annotated Corpus of Verbal MWEs for English. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 193–200, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2020. Annotating verbal MWEs in Irish for the PARSEME Shared Task 1.2. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, MWE-LEX@COLING 2020, online, December 13, 2020*, pages 58–65. Association for Computational Linguistics.