Negative-polarity Multiword Expressions (NPMWEs)
Interpreting Corpus Results and Enriching a Multilingual Resource

Monica-Mihaela Rizea*, Gianina Iordăchioaia[+], Frank Richter[°]
* Solomon Marcus Center for Computational Linguistics, Bucharest, [+] University of Stuttgart, ° Goethe-University, Frankfurt a.M.

[WG1][1]

This poster is dedicated to the analysis of a special category of multiword expressions, negative-polarity MWEs, with the aim of enriching the Romanian Collection of Negative Polarity Items (CoDII-NPI.ro), which is part of a multilingual resource, CoDII, hosted by Institut für England und Amerikastudien (IEAS)[2]. The distribution of this type of MWEs, as members of the larger class of Negative Polarity Items (NPIs), is generally considered to be restricted to certain licensing contexts, prototypically negative or negative-like environments (such as interrogatives, antecedents of conditionals, or modifiers of superlative and universal NPs), even if they do not, themselves, express negation. Typical NPI examples are lexical items such as *any*, *ever* but also multiword expressions such as (*say) a word*. A known fact is that some of these expressions, for reasons as yet unclear, have a more restricted occurrence pattern than others and exhibit an idiosyncratic behaviour in relation with their licensers. Negative-polarity Multiword Expressions (NPMWEs) are analysed here as collocationally restricted lexical units i.e. units that display a collocate-collocator relation with their licensing contexts[3]. The collocational, representational account allows determining statistical profiles for NPMWEs from corpora and a classification according to their distributional dependence on the licensing contexts. This information is then included in a growing multilingual database which facilitates modelling the idiosyncratic variation for NPI expressions both at the level of individual languages and from a cross-linguistic, comparative approach.

Strategies for enhancing the Romanian NPI database as part of a multilingual resource

Step 1 Collecting the items (paradigmatic level). For the purpose of this study, a collection of five Romanian general dictionaries was used: these dictionaries are accessible via an online database[4] that also allows queries using regular expressions and can generate results from the text of the glosses. The definitions provide usage information such as "especially in negative contexts". After automatically generating a list of candidates, a Romanian linguist selected 100 NPMWE candidates for further analysis. The original Romanian NPI database contained 58 items, only one layer of syntactic information and no corpus examples. Currently, context examples are being included for every type of licenser and the syntactic information is updated.

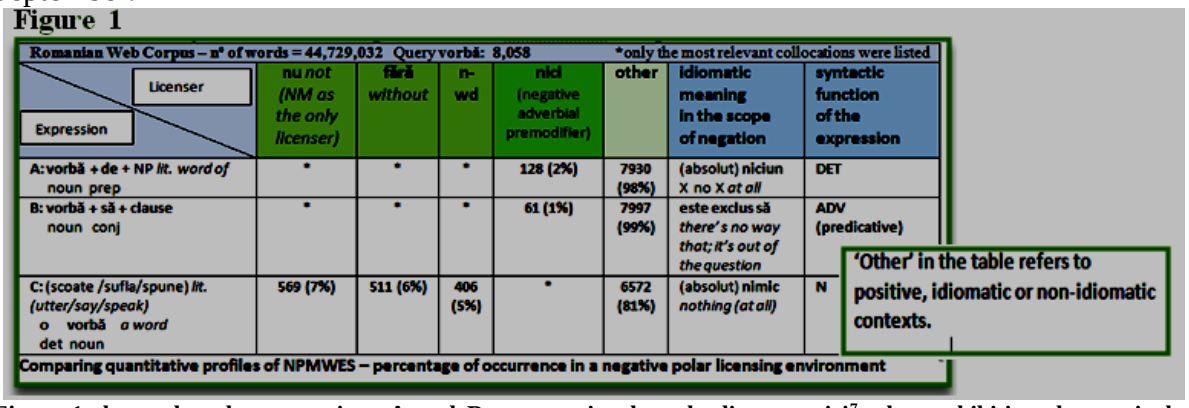
Step 2 Analysis of contextual profiles (syntagmatic level). Each item is then investigated in terms of occurrence patterns and relevant context examples in order to document the compatibility with each category of licenser using the *sketchengine.co.uk* tool[5] (we collect the results in a table as in Figure 1). The corpus profile reveals the licensers for an NPMWE. We use this information to classify an NPMWE as superstrong, strong or weak.[6] The process of integrating each unit in the NPI database implies judgments in terms of **a.** distinguishing between different expressions that share a common element and that show idiomatic meaning(s) when used in the scope of negation **b.** identifying the idiomatic meaning(s) of MWEs that are used in the scope of negation and separating polysemous units **c.** completing

[1] Discussions with Manfred Sailer during my STSM in Frankfurt were essential in determining a number of important decisions on modifying and extending the current NPI collection.
[2] http://www.english-linguistics.de/codii/.
[3] This is in line with previous analyses developed by Soehn et al. 2010, following van der Wouden 1997.
[4] https://dexonline.ro/.
[5] Info about the corpora hosted and tool functions is available at: https://www.sketchengine.co.uk/.
[6] For theoretical background, see van der Wouden 1997.

syntactic information on two levels: *syntactic characterization of the parts* and *syntactic function of the expression* **d.** identifying compatibilities with different licensers and providing corpus examples to document each combination.

State of completion of the work. By the time of submission, 100 NPMWE candidates have been extracted. 20 of them have been added to the current database (which is an extension of 35%). We also expect to have the corpus profiles for the 20 newly added NPMWEs by September.

Figure 1



Romanian Web Corpus – n° of words = 44,729,032 Query vorbă: 8,058 *only the most relevant collocations were listed

Licenser / Expression	nu *not (NM as the only licenser)*	fără *without*	n- wd	nici (negative adverbial premodifier)	other	idiomatic meaning in the scope of negation	syntactic function of the expression
A: vorbă + de + NP *lit. word of* noun prep	*	*	*	128 (2%)	7930 (98%)	(absolut) niciun X no X *at all*	DET
B: vorbă + să + clause noun conj	*	*	*	61 (1%)	7997 (99%)	este exclus să *there's no way that; it's out of the question*	ADV (predicative)
C: (scoate /sufla/spune) *lit. (utter/say/speak)* o vorbă *a word* det noun	569 (7%)	511 (6%)	406 (5%)	*	6572 (81%)	(absolut) nimic *nothing (at all)*	N

'Other' in the table refers to positive, idiomatic or non-idiomatic contexts.

Comparing quantitative profiles of NPMWES – percentage of occurrence in a negative polar licensing environment

***Figure 1* shows that the expressions A and B are restricted to the licenser nici[7] when exhibiting the particular idiomatic meaning in the scope of negation. The expression C is compatible with a wider range of negative contexts.**

Corpus examples:

(1) [...] când e nevoie de asistenţă medicală, ***nici* vorbă de** medic.
when is need of assistance medical no word of doctor
*when medical assistance is needed, there's **no** doctor **at all.***

(2) ***Nici* vorbă să** fie spion [...]
no word SJ be spy
***it's out of the question** that he's a spy [...]*

(3) Brusc îşi dădu seama că ***nimeni*** nu scotea[8] **o vorbă** *[...]*
suddenly has figured that nobody NM utter a word
*He shortly figured that **nobody** said **anything at all.***

(4) Stau chiar aşa, în faţa cafenelei, ***fără*** să scoată **o vorbă**?
stay like this in front of café.the without SJ utter a word
*So are they staying like this, in front of the café, **whithout** saying **anything at all**?*

Relevance to PARSEME and to WG1. NPMWEs are a theoretically and practically challenging class of multiword expressions because their obligatory licensers are not simple lexemes but abstract grammatical and semantic categories. Linguistic documentation (in terms of syntactic, semantico-pragmatic and contextual information) of Negative Polarity MWEs offers rich information for annotation tasks and for experiments of automatic extraction.

REFERENCES

Hoeksema, Jack. 2009. Jespersen recycled. In Cyclical change, ed. Elly van Gelderen, 15–34. Amsterdam/Philadelphia, Benjamins.

Ton van der Wouden. 1997. Negative Contexts. Collocation, polarity and multiple negation. Routledge, London and New York.

Jan-Philipp Söhn, Beata Trawiński, Timm Lichte. 2010. Spotting, collecting and documenting Negative Polarity Items. Natural Language and Linguistic Theory, 28, 931–952.

Van der Wal, Sjoukje. 1996. Negative polarity items and negation: Tandem acquisition. Doctoral Dissertation, University of Groningen.

[7] For a similar discussion of 'idiomatized NPIs' or NPIs licensed in only one environment, see van der Wal 1996.

[8] Similarly to the English counterpart, the predicates combining with the NPI 'o vorbă' display some variation (see Hoeksema 2009 for a detailed description), which makes 'a scoate' not a lexicalized part of the expression.