A Corpus-based Analysis of Negative Polarity Idiomatic Expressions

Monica-Mihaela Rizea[1], Gianina Iordăchioaia[2], and Frank Richter[3]

[1]Solomon Marcus Center for Computational Linguistics, University of Bucharest, monicamihaelarizea@gmail.com
[2]University of Stuttgart, gianina-nicoleta.iordachioaia@ling.uni-stuttgart.de
[3]Goethe University, Frankfurt a.M., f.richter@em.uni-frankfurt.de

This abstract presents some results of an analysis dedicated Negative Polarity Idiomatic Expressions, following the collocational approach to NPI licensing (Richter and Soehn 2006; Sailer 2009). Corpus-linguistic methods were applied in order to determine individual statistical profiles and a classification (into superstrong/strong/weak - van der Wouden 1997) of these expressions according to their distributional dependence on licensing contexts. The study is correlated with a practical task: updating the sub-collection of Romanian Negative Polarity Items (CODII.NPI.ro)[1] and enriching it with Negative-polarity Multiword Expressions (NPMWEs), which are understood as collocationally-restricted lexical units with idiosyncratic distributional patterns. After generating a list of NPMWE candidates from the existing lexicographic resources (Romanian general dictionaries accessible via an online database (dexonline.ro) and *The Dictionary of Romanian Expressions, Syntagms and Phrases* – DELS 2010), some of them were selected for further investigation. The NPMWE candidates were checked against the Romanian Web Corpus (- ROWac - n^{o} of words = 44,729,032) via the Sketch Engine online tool[2] by analysing their individual realizations in context, in relation to a predefined set of licensers[3]. One challenge that is associated with the corpus analysis when trying to determine statistical profiles for every licenser-NPMWE collocation implies the ability to discriminate between "competing" MWE structures that appear in corpora. This implies cases such as[4]: **1.** NPMWEs that have *common lexemes*, but different idiomatic readings (and, generally, different syntactic structures) **2.** NPMWEs that have common lexemes, same idiomatic meaning and same syntactic structure **3.** NPMWEs that have *common lexemes* with non-negative polar expressions **4.** formally *identical* MWE structures that exhibit constructional "layering" (see Hoeksema 1994), having several uses, some of which are negative sensitive while others are not. Therefore, a first step in the investigation of a NPMWE candidate is disambiguation; for this purpose, we built lists (extracted from the lexicographic resources) of related expressions in order to take them as reference points in the analysis[5].

SOME EXAMPLES (considering both the original collection and the new NPMWE list)

In the Romanian NPI collection, there are currently two NPMWEs listed under the entry "vedea" ("see") – **1.** "(de/că nu) se vedea om cu om" (gl. "(that not) each other see person with person" – part of a

[1]This collection is part of a (comparable) multilingual electronic resource (CODII) in XML format, hosting German, English and Romanian collections of distributionally idiosyncratic items (www.english-linguistics.de).
[2] the.sketchengine.co.uk
[3] These licensers are listed in the CODII.NPI.ro Licensing Contexts section and acquire binary ("yes"/"no") values according to the specific distributional patterns of each NPI entry.
[4] Some examples are provided below.
[5] Even if the corpus itself brings useful information about alternative MWE structures (i.e. the syntagmatic analysis), we also take into account the paradigmatic level.

result clause – lit. "that two persons could not see each other") and **2.** "(ce/cum n-) a văzut Parisul" (gl. "(that not=)have seen Paris" – part of a relative clause – lit. "that Paris has not seen"). These expressions have the verb "see" as a common element and they are both minimizers in Romanian (even if, as part of result/relative structures, they appear with a reversed, intensifying function in context[6]); however, they cannot be used interchangeably and have different syntactic structures, which motivated the decision of treating them as separate entries:

NPMWE 1 se vedea om cu om (PRON VERB NOUN ADP NOUN) (the less likely alternative - not even seeing the other person in your range of sight)

This structure has two usage patterns, with the first one being more semantically motivated:

(1) [...] era şi întuneric şi **ningea de nu se vedea om cu om**. (ROWac)
was even dark and snowing that not CL see person with person
[...] it was dark and *snowing* so abundantly that *you couldn't see an inch in front of you*.

(2) E acolo o **mafie de nu se vede om cu om.** (http://zdbc.ro/patronul-de-la-nevila-amendat-cu-100-de-milioane-de-catre-primaria-onesti/)
is there a mafia that not CL see person with person
The *mafia* there is *unbelievable*. (pointing to incredibly high rates of criminal activity)

NPMWE 2 a văzut Parisul (AUX VERB NOUN) (the less likely alternative – not even Paris (metonymy - city for its inhabitants) has seen this); *pejorative*

(3) Pentru alţii, această **ţigănie ce n-a văzut Parisul** este prilej de acută melancolie… (http://www.contributors.ro/societatelife/tiganie-ce-n-a-vazut-parisul%E2%80%A6/)
For others, this *unprecedented gypsyism* is the trigger of an acute melancholy...

As the corpus analysis suggests, both NPMWEs have **'not'** as the only licenser, showing a very restricted occurence pattern (which classifies them as superstrong).

Additionally, the verb "see" is also a common element of other MWEs from the candidate list such as "vedea(/privi/se uita) cu ochi buni" (lit. "see(/watch/look) with good eyes"; idiomatic meaning in the scope of negation: *to disapprove, to disallow; to dislike*) or "vedea pe cineva bine" (lit. "see somebody well"; idiomatic meaning in the scope of negation: *to foresee (that somebody will get into) trouble (as prevention or threat*)), etc.; these expressions prove to be compatible with a wider range of licensing contexts, including downward-entailing operators such as few (which classifies them as weak)[7]. So far, the analysis illustrates the **case 1** described above.

If we take a look at the expressions "vedea(/privi/se uita) cu ochi buni" (currently, only "privi cu ochi buni" is registered in the database, and with a positive meaning), we notice that they show different percentages in terms of occurrence in positive vs. negative contexts, but with a clear preference for negative environments. "Vedea cu ochi buni" appears 43 times in ROWac, with 93% preference for negative polarity contexts, "privi cu ochi buni" appears 29 times, with 86% occurrence in negative-like environments, while "se uita cu ochi buni" is attested 7 times in the corpus, with same 86% preference for negative contexts. This would be an illustration of **cases 2** (common lexemes, same idiomatic meaning and same syntactic structure – VERB ADP NOUN ADJ) and **4** since each expression has also a positive use, that of "to look kindly on somebody/something, "to think well of somebody/something", "to approve", even if lowly attested in corpora. **Case 3** is exemplified by the NPMWE "închide **un ochi**"

[6] E.g.: If the consequence of an event such as snowing in (1) is the lowest scale of visibility, then the degree of "snowing" is highest (the relevant minimal values lead to maximal interpretations in context).

[7] We will not describe the analysis at this point.

(lit. ”close one eye”; with an idiomatic meaning in the scope of negation similar to the English minimizer expression ‘sleep a wink’ (i.e. not to sleep at all)) in contrast with resembling (polarity neutral) MWEs such as “închide **ochii**” (unlike the first structure, the noun has an obligatory definite article; lit. close the eyes; idiomatic meaning – *die*) or “arunca un ochi” (lit. throw an eye – idiomatic meaning: *to inspect something quickly*).

CONCLUSIONS

Corpus investigation offers the perspective of “competing” structures in use, which refines the analysis of NPMWEs by taking into account different cases of ambiguity and documenting them at a theoretical level. Moreover, displaying this information in the CODII.NPI.ro database (in XML format) can also be used a (disambiguation) resource in experimental tasks of automatic extraction.

REFERENCES

DELS. 2010. Dicționar de expresii, locuțiuni și sintagme ale limbii române ('The Dictionary of Romanian Expressions, Syntagms and Phrases'), Cătălina Mărănduc. Bucharest: Corint.

Hoeksema, Jack. 1994. On the grammaticalization of negative polarity items, in Gahl, S. et al., (eds.), Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistic Society, Berkeley, pp. 273-282.

Richter, Frank and Jan-Philipp Soehn. 2006. Braucht niemanden zu scheren: A Survey of NPI Licensing in German. In S. Muller (Ed.), The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar, Stanford: CSLI Publications, pp. 421-440.

Sailer, Manfred. 2009. A representational theory of negative polarity item licensing. Habilitation thesis, Universität Göttingen.

Soehn, Jan-Philipp, Liu, Mingya, Trawiński, Beata, and Gianina Iordachioaia. 2010. Nicht sonderlich oder doch sattsam bekannt? Positive und Negative Polaritätselemente als lexikalische Einheiten mit Distributionsidiosynkrasien EUROPHRAS 2008. Helsinki, pp. 273-281.

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