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

Monica-Mihaela Rizea[1]; Gianina Iordăchioaia[2]; Frank Richter[3]
[1]Solomon Marcus Center for Computational Linguistics, [2]University of Stuttgart, [3]Goethe University, Frankfurt a.M.

PARSEME 7th General Meeting, Dubrovnik,26-27 September 2016

NPMWEs

Negative-polarity Multiword Expressions are distributionally restricted lexical units that occur in the scope of negation, but also in a variety of other contexts such as questions, the scope of conditional operators and of downward-entailing expressions, in the restrictor of universal quantifiers, the scope of imperatives, etc. They display non-referential, idiomatic meanings that are specific to the negative-like environments in which they appear, and that are only possible as long as their distributional restrictions are respected.

(1) **N**-am văzut **picior de** *om* în pădure.
not=have seen leg of person in forest
I haven't seen **a living soul** in the forest [= I haven't seen **anyone at all** in the forest.]
(2) **#**Am văzut **picior de** *om* în pădure.
have seen leg of person in forest
I have seen a living soul in the forest.

MWEs that are sensitive to negative polarity are a problematic class both for traditional lexicography and for computational applications since their obligatory licensers are not simple lexemes, but abstract grammatical and semantic categories (Trawiński et al. 2008).

Example

GLOSS ***Ţipenie de*** *("shout.suffix of" - lit. "no one to shout", "no living creature")* is part of a complex nominal phrase of the type *N1 + DE + N2*, i.e. *"ţipenie" + "de" + N2* (in the context of a licenser), which, as a whole, functions as an emphatic negator (Dindelegan 2013:128). N2 has limited lexical variation, usually reduced to ***"om" ("human"/"person").*** In the **Romanian Web Corpus (roWaC)**, there are also other realizations of N2, such as *"vietate" ("creature")* and *"terorist" ("terrorist")*. An English correspondent can be found in the minimizer construction **"a (living) soul"**. Similarly to the English expression, "ţipenie de" has the **idiomatic meaning** *"anyone at all"/"absolutely anybody"* in negative contexts. "Ţipenie" usually occurs as a bare noun when preceded by the scalar negator *"nici" ("not even")*; however, it can also be preceded by the negative determiner ***"nicio"*** *("no.fem")*. There are also contexts where "ţipenie de" can appear with **clausemate negation** and no other negative element. This expression is only felicitous in negative contexts. It is part of a special class of Romanian **minimizers** such as *"urmă/umbră/suflare/suflet de"* – etymologically, *"trace/shadow/breath/soul of"* that combine with non-gradable entities, usually [+animate] (except for "urmă" that also occurs with other types of non-scalar N2) and that can be considered as the faintest manifestations of N2 on a scale of perception. This is a valid mechanism for obtaining emphatic NPI minimizers: negating the minimum imaginable evidence of the existence of an entity N2 rendered by something that is not even a part of N2, not even a material attribute of the entity it stands for. Just like the other minimizers, they evoke the least likely alternative to the entity in focus, which is, actually, N2, the semantic head of the structure. Since N2 has a very limited lexical variation, "ţipenie" is many times used alone and it incorporates the meaning of N2: *"Ziua nu întâlneau ţipenie."* (*lit. "During daytime, they didn't meet living.soul"*), meaning *"they didn't meet anyone at all."* For example, in **roWaC,** from **114** occurrences of the word "ţipenie", **30** occurrences (i.e. **26%**) represent cases when "ţipenie" is used without N2.

CORPUS EXAMPLES (source : roWaC)

CMN (sentential negation - NM **nu** "not")
(3) **Nu** se zărea **ţipenie de** *om*, locul părea pustiu.
not CL saw ŢIPENIE of person place.the seemed deserted
Not a living soul in sight, the place seemed deserted [= There **wasn't anyone** in sight, the place seemed deserted.]

NW (negative determiner **nicio** "no")
(4) Pe drum, **nicio ţipenie de** *om*.
on road, no ŢIPENIE of person
On the road, **no living soul.** [= There wasn't **anyone at all** on the road.]

NICI (scalar negator **nici** "not.even")
(5) Poate de aceea nu e **nici** **ţipenie de** *om* în jur.
Maybe that.is.why not is not.even ŢIPENIE of person around
Maybe that's why there's **no living soul** around. [= Maybe that's why there **isn't anyone** around.]

WHITHOUT ("fără")
(6) Am trecut prin pădure, spre calea ferată, **fără** să întâlnim **ţipenie de** *om*.
have passed through forest to railway.the without SJ meet ŢIPENIE of person
We passed through the forest, to the railway, **without** meeting **a living soul.** [= We passed through the forest [...], **without** meeting **anyone at all.**]

READING-DEPENDENT LICENSING (source: linguist)

DENT (downward-entailing operator *puţini/puţine* "few")
(7) **#Puţini** călători au întâlnit **ţipenie de** *om* în deşert.
few travelers have met ŢIPENIE of person in desert
Few travelers met **a living soul** in the desert.

NCMN (negated verbs – *pretinde* "claim"/*crede* "think")
(8) **#Nu pretind** că am văzut **ţipenie de** *om* în noaptea aceea.
not pretend that have seen ŢIPENIE of person in night.the that
I don't claim I've seen **a living soul** that night.
(9) **Nu cred** c-am văzut **ţipenie de** *om* în noaptea aceea.
not think that=have seen ŢIPENIE of person in night.the that
I don't think I've seen **a living soul** that night.

NV (inherently negative matrix verbs such as *a fi surprins(ă)* "be surprised" or *a se îndoi* "doubt")
(10) **#Mă surprinde** că văd **ţipenie de** *om* în deşert.
CL surprise that see ŢIPENIE of person in desert
I'm surprised that I see **a living soul** in the desert.
(11) **Mă îndoiesc** că voi vedea **ţipenie de** *om* în deşert.
CL doubt that will see ŢIPENIE of person in desert
I doubt that I'm going to see **a living soul** in the desert.

QUE (in negatively biased rhetorical questions)
(12) Speri să întâlneşti **ţipenie de** *om* pe drum la ora asta?
hope SJ meet ŢIPENIE of person on road at hour this
Do you hope to meet **a living soul** on the road at this hour?

IF (in conditional threats, episodic statements, conditional promises)
(13) Dacă văd **ţipenie de** *om* în această rezervaţie naturală, îmi voi ieşi din minţi! (threat reading)
If I see **a living soul** in this protected nature area, I will go mad!
(14) **#**Dacă întâlnesc **ţipenie de** *om* în deşert, îl salut. (episodic reading)
If I meet **a living soul** in the desert, I say hello.
(15)(#) Dacă întâlnesc **ţipenie de** *om* în noaptea asta în bar, plătesc toată băutura. (promise reading)
If I meet **a living soul** at the bar tonight, I'll pay for all the drinks.

Comment: - **(15)** is infelicitous with a literal reading; felicitous with an "affirmative sarcasm" reading - the condition for a promise can be paraphrased as **"I strongly doubt that...";** a case of sarcasm licensing (Horn 2016) - the interpretation is negative as a sarcastic effect.

(16) Dacă întâlneşti **ţipenie de** *om* în noaptea asta, **chiar** eşti norocos. (the condition has a reading **"I strongly doubt"**)
If you meet **a living soul** tonight, you're **really** lucky. (->I strongly doubt there's any chance for you to meet another person tonight).

UNIV (in the restrictor of a universal quantifier)
(17) **#**Oricine întâlneşte **ţipenie de** *om* în deşert, îl salută. (episodic reading)
Everyone who sees **a living soul** in the desert, says hello.
(18) Oricine vede **ţipenie de** *om* în bezna asta, se poate considera binecuvântat. ("I strongly doubt" reading)
Whoever sees **a living soul** in this darkness can consider him/herself blessed.

CONCLUSION

This expression behaves like a **strong** NPI (see Sailer 2009 a & b for details on reading-dependent licensing of strong NPIs) since it is not licensed by DE determiners such as **"puţini/puţine" ("few")**, it is felicitous in the context of **neg. raising** predicates such as **"nu cred"("I don't think")**, but it is strange in the complement clause of **"nu pretind" ("I don't claim");** it is licensed by **negative predicates** such as **"a se îndoi" ("doubt"),** but not by **"a fi surprins(ă)" ("be surprised");** "ţipenie de" is licensed in **rhetorical questions** and in the **antecedent of conditional threats** but it is not licensed in episodic readings; it is not felicitous with conditional promises, unless they receive a **sarcastic interpretation** (the condition for a promise is interpreted as **"I strongly doubt that..."**); it is licensed in the restrictor of a universal quantifier only if the context receives the **"I strongly doubt that..."** interpretation.

The importance of documenting NPMWEs

CODII.NPI.ro is a freely available electronic resource that lists and documents (in terms of syntactic, semantico-pragmatic, contextual info, and English translations) the Romanian expressions that have an idiomatic meaning specific to negative contexts. Documenting NPMWEs for multiple languages does not only facilitate comparative linguistic studies, but it could also represent a useful resource for translators that search for paraphrases of idioms sensitive to negative polarity and also for second language learners who can find real-use examples from corpora (with English glosses and translations) for every licenser-NPMWE pair. Moreover, the XML database offers rich information for applied tasks: corpora annotation, parser training, experiments of automatic extraction of (NP)MWEs, etc. and can be exploited by NLP applications such as electronic dictionaries, MT or CAT software.

Objectives

PRACTICAL TASKS

- Update the initial entries listed in the Romanian Collection of Negative Polarity Items (CODII-NPI.ro) - 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).
- Enrich CODII-NPI.ro with NPMWEs.

THEORETICAL TASKS

- Classify the (new) Romanian NPMWEs into superstrong/strong/weak.

CRITERIA (following van der Wouden 1997):

- Superstrong NPIs are licensed only by antimorphic contexts (overt negation).
- Strong NPIs are licensed by antimorphic and anti-additive (comprising n-words and without) contexts.
- Weak NPIs are licensed by antimorphic, anti-additive, and downward-entailing contexts (plus the remaining ones).

We further refined these criteria by considering reading-dependent licensing in the case of strong NPIs (Sailer 2009 a & b).

FRAMEWORK

- The collocational account allows us to determine corpus profiles in terms of licenser-NPMWE collocations (the distributional dependence on the licensing contexts is documented with frequency data and real-use examples from large Romanian corpora such as roWaC (nº of words = 44,729,032) or OPUS2 Romanian parallel corpus (nº of words = 282,408,295) via the Sketch Engine online tool (the.sketchengine.co.uk)).
- In this approach, NPMWEs are understood as collocationally-restricted lexical units with idiosyncratic distributional patterns .

Methodology

STEP 1 Collecting the Items (Paradigmatic Level)

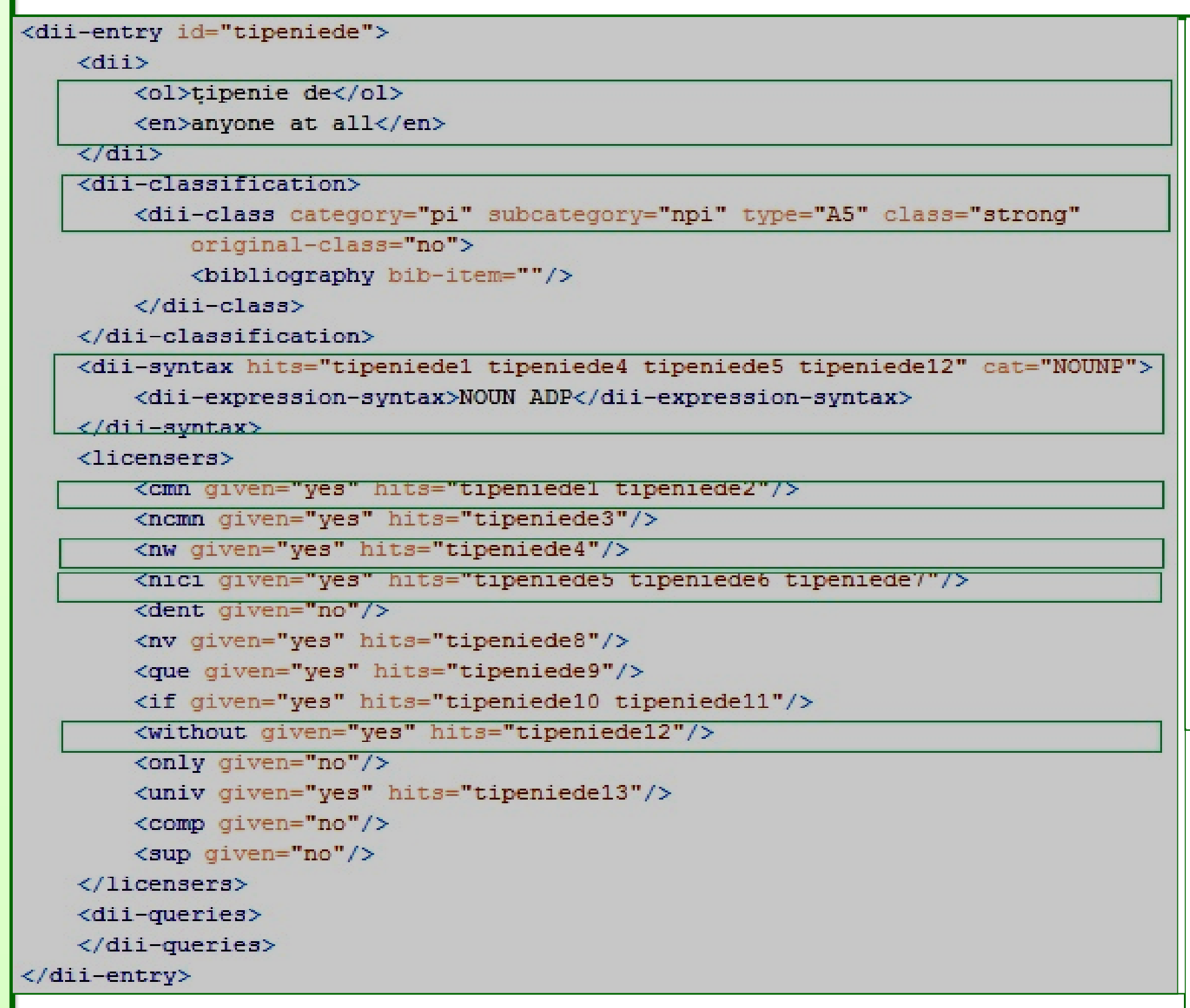
After generating a list of candidates from the existing lexicographic resources (that do not place special focus on negative polarity), 100 NPMWE candidates were selected for further analysis. The dictionaries are accessible via an online database (dexonline.ro) that also allows queries using regular expressions and can generate results from the text of the glosses. Additional resource: DELS 2010 (*The Dictionary of Romanian Expressions, Syntagms and Phrases*)

NOTE: Only in extremely rare occasions do the definitions provide usage information such as ("in negative constructions/sentences") (e.g. only 48 times in DEX - *The Explanatory Dictionary of the Romanian Language with* ≈ 67 000 entries). The only other possibility to find such expressions in the dictionary is when they are listed with a negative element such as "nu" ("not") or "nici" ("not even"). For example, from 11430 expressions listed in DELS, 518 (i.e. 4.5%) contain "nu" and/or "nici" (but these expressions are not necessarily NPIs).

STEP 2 Analysis of Contextual Profiles (Syntagmatic Level)

We investigate the candidates in terms of occurrence patterns and real-use (corpus) examples in order to document the compatibility with each category of licensers.

TOOL: Sketch Engine; no corpus examples - use linguist intuition


```
<dii-entry id="tipeniede">
    <dii>
        <ol>ţipenie de</ol>
        <en>anyone at all</en>
    </dii>
    <dii-classification>
        <dii-class category="pi" subcategory="npi" type="A5" class="strong"
            original-class="no">
            <bibliography bib-item=""/>
        </dii-class>
    </dii-classification>
    <dii-syntax hits="tipeniede1 tipeniede4 tipeniede5 tipeniede12" cat="NOUNP">
        <dii-expression-syntax>NOUN ADP</dii-expression-syntax>
    </dii-syntax>
    <licensers>
        <cmn given="yes" hits="tipeniede1 tipeniede2"/>
        <ncmn given="yes" hits="tipeniede3"/>
        <nw given="yes" hits="tipeniede4"/>
        <nici given="yes" hits="tipeniede5 tipeniede6 tipeniede7"/>
        <dent given="no"/>
        <nv given="yes" hits="tipeniede8"/>
        <que given="yes" hits="tipeniede9"/>
        <if given="yes" hits="tipeniede10 tipeniede11"/>
        <without given="yes" hits="tipeniede12"/>
        <only given="no"/>
        <univ given="yes" hits="tipeniede13"/>
        <comp given="no"/>
        <sup given="no"/>
    </licensers>
    <dii-queries>
    </dii-queries>
</dii-entry>
```
FIG. 1. CODII.NPI.ro - XML representation of the Licensing Contexts section

Romanian Web Corpus – nº of words = 44,729,032 **Query:** *ţipenie de om 'a living soul'* N = 81	
negative(-like) (A+B+C+D+E+F)	N= 76 (94%)
A:nu (n-/ne-) 'not' NM as the only licenser	N= 34 (42%)
B:nw	N= 2 (2%)
C:nici 'not even'	N= 37 (46%)
D:fără 'without'	N= 3 (4%)
E: NV	N= 0 (0%)
F: dacă 'if'	N= 0 (0%)
Other	N= 5 (6%)

Table 1 . Corpus Profile – Distribution of *ţipenie de om* with respect to its licensers

Romanian Web Corpus – nº of words = 44,729,032 **Query:** *picior de om 'a living soul'* N = 50	
negative(-like) (A+B+C+D+E+F)	N= 43 (86%)
A:nu (n-/ne) 'not' NM as the only licenser	N= 38 (76%)
B:nw	N= 1 (2%)
C:nici 'not even'	N= 4 (8%)
D:fără 'without'	N= 0 (0%)
E: NV	N= 0 (0%)
F: dacă 'if'	N= 0 (0%)
Other (positive uses including the non-idiomatic reading of a 'body part')	N= 7 (14%)

Table 2 . Corpus Profile – Distribution of *picior de om* with respect to its licensers

OPUS2 English – nº of words = 1,139,515,048 **Query:** *'a living soul'* N = 134	
negative(-like) (A+B+C+D+E)	N= 103 (77%)
A: not (n't) NM as the only licenser	N= 69 (51%)
B: nw	N= 34(25%)
C: without	N= 1 (1%)
D: NV	N= 0 (0%)
E: if	N= 0 (0%)
Other (positive uses including the 'walking dead' meaning)	N= 31 (23%)

Table 3 . Corpus Profile – Distribution of *a living soul* with respect to its licensers

Comparative corpus profile representations of two synonymous negative-polarity MWEs (*picior de* "leg of" and *ţipenie de* "ŢIPENIE of" collocating with *om* "person" and (one) of their English equivalent NPMWEs *a living soul*).

Improvements (initial sample-20 entries)

1. Syntactic information to include the characterization of the individual parts of the expression and of the entry as a whole (POS tags used: http://universaldependencies.org/ro/pos/index.html.) 2. Corpus examples (from large Romanian corpora - such as roWaC or OPUS2 parallel corpora - via the Sketch Engine online tool) for every licenser-NPMWE pair 3. Statistical profiles for every licenser-NPMWE collocation as they are reflected in roWaC 4. Subcorpora generated from filtering specific licenser-NPMWE contexts (information that will be used for further theoretical and applied studies) 5. Information about 'competing MWEs' (including cases of polysemy when the expressions might also exhibit non-NPI senses) so as to avoid ambiguity (both for human readers and for tasks of automatic extraction) 6. In the case of strong NPMWEs, we also document reading-dependent licensing cases (some categories of licensers only license strong NPIs under specific readings - see Sailer 2009 a & b for details).

References

DELS. 2010. Dicţionar de expresii, locuţiuni şi sintagme ale limbii române ('The Dictionary of Romanian Expressions, Syntagms and Phrases'), Cătălina Mărănduc. Bucharest: Corint.
Dindelegan, G. P., The Grammar of Romanian, Oxford University Press, 2013.
Horn, L.R. 2016. Licensing NPIs: Some Negative (and Positive) Results. In P. Larrivée & C. Lee (Eds.). Negation and polarity: Experimental perspectives, Cham: Springer, pp. 281-305.
Ionescu, Emil. 2004. Studies on Negation in Romanian: Past and Present. In E. Ionescu (Ed.). Understanding Romanian Negation. Bucharest University Press, pp. 13-31.
Losnegaard, G., Sangati, F., Parra, C., Savary, A., Bargmann, S., and J. Monti. 2016. PARSEME Survey on MWE Resources, Proceedings of LREC'16.
Sailer, Manfred. 2009 a. On reading-dependent licensing of strong NPIs. In A. Riester and T. Solstad (Eds.). Proceedings of Sinn und Bedeutung 13, Stuttgart, pp. 455-468.
Sailer, Manfred. 2009 b. A representational theory of negative polarity item licensing. Habilitation thesis, Universität Göttingen.
Soehn, Jan-Philipp, Trawiński, Beata, and Timm Lichte. 2010. Spotting, collecting and documenting Negative Polarity Items. Natural Language and Linguistic Theory, 28, pp. 931–952.
Trawiński, Beata, Soehn, Jan-Philipp, Sailer, Manfred, and Frank Richter. 2008. A Multilingual Electronic Database of Distributionally Idiosyncratic Lexical Items. Proceedings of Euralex 2008.
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.

This work was initiated during a Short Term Scientific Mission at the Institute for English and American Studies (IEAS), Goethe University, Frankfurt a.M.