Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет
«Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа
«Фундаментальная и компьютерная лингвистика»

Бибаева Мария Александровка

Автоматическое извлечение единиц отрицательной полярности

Automatic extraction of polarity sensitive items

Выпускная квалификационная работа студента 4 курса бакалавриата

Академический руководитель образовательной программы	Научный руководитель
канд. филологических наук, доц.	Доктор наук, доцент
Ю.А.Ландер	Ф.Тайерс
……2018 г.	Научный консультант Доктор филологических наук, профессор В.Ю.Апресян

Москва 2018

Contents

1 Introduction

Negative and positive polarity items (NPIs and PPIs) have been the subject of research in semantics, formal semantics, syntax, pragmatics and to some extent typology since Klima's survey of negation in English. The main idea is that NPIs are items which have distribution limited to a specific set of contexts. Naturally, every word's distribution is limited somehow, but in case of NPIs these limitations form a special pattern, in which the distribution of them is limited mostly to environments that count as negative in some way. The classic example is the English indefinite pronoun *any*. The example (1), taken from both (Ladusaw 1979) and (Wouden T. 1994), demonstrates, that sentences containing *any*-pronouns can be qualified as grammatical only if they also have some negative marker.

(1) a. *John has talked to any of the students.
    b. *He did anything to help her.

(2) a. John has**n't** talked to any of the students.
    b. He did**n't** do anything to help her.

Examples (1) and (2) show only a part of *any*'s distribution as it is not limited to direct negation only. There are some other contexts that can *license* polarity items. While an exhaustive set of all relevant licensing contexts and their analysis is not yet agreed upon, a lot of progress has been already made. The thing is that all contexts which can license NPIs must have some common feature in common, and it is this feature that is debated. By now, there are several widely accepted approaches which will be discussed in Section 2.

Unfortunately, the majority of literature on NPIs describes a rather small set of items which are mostly indefinite pronouns and other frequent items such as modal verbs and several adverbs and idioms. As can be demonstrated by examples like 1, polarity sensitivity is a feature not only of these types of units. The reason for such neglect is that it is plain easier to find an NPI among pronouns or modal verbs than among lexical verbs or adjectives, since grammatical meanings which an NPI can have are fairly close and predictable. The same is true for idioms with the meaning of *small amount or activity*, such as (3), which are regularly described as NPIS.

(3) a. He did**n't** lift a finger to help her.
    b. *He lifted a finger to help her.

That is why to make any progress in this area of research would require some new systematic way of collecting data that would enable to expand the range of polarity items being considered. In this work we tried to make such kind of system, extracted polarity sensitive verbs from Russian corpora and compared

our result to the result of (Апресян 2017) where similar research was conducted manually.

2 Theoretical background

As was shown in Section 1, the NPIs are items with restricted disctribution, or, as (Wouden T. 1994) defines them:

(4) *NPIs are expressions that can only appear felicitously in negative contexts.*

However, this particular definition is not very precise. As mentioned earlier, NPIs can appear not only in the scope of direct negation, but in several other contexts which are called *licensing contexts*. Moreover, a negative context can be constituted by lexical elements of almost any category

(5) *Verbs*
We want to **avoid** doing any lookup, if possible.

(6) *Adverbs*
   a. Many people sympathized with my viewpoint, but - I repeat - **hardly** any practical solutions were given.
   b. I've **never** come across anyone quite as brainwashed as your student.

(7) *Nouns*
   a. The positive degree is expressed by the **absence** of any phonic segment.
   b. He is right about the **lack** of anything resembling sound correspondences in syntax.

(8) *Prepositions*
You can exchange **without** any problem/obligation.

Not only particular words can constitute a negative context but whole constructions such are

(9) *Questions*
We want to **avoid** doing any lookup, if possible.

Which are the licensing conditions of polarity items is one of the major questions in this area and the base of all research methods at the same time. Every paper on polarity sensitive items, computational or theoretical, takes the licensing contexts as a scale of polarization. However, an exhaustive set of relevant feature has not yet been established. Two major approaches to licensers which have been showing the best results are described in Sections 2.1 and 2.2.

However, what can be said about the general use of NPIs is that sentences containing them are not syntactically incorrect if we delete a licensing marker from them. What makes a sentence with an NPI in absence of a licenser ungrammatical is the troubling semantics of an NPI. This is what Ladusaw defined as *Semantic Filtering*:

(10) *Semantic filtering*
grammatical = def **Syn** (ϕ) $\wedge$ **Sem** (ϕ);
where **Syn** is syntactic well-formedness, **Sem** is semantic well-formedness

So, a sentence with an NPI must be not only syntactically correct but also semantically correct which means that a presence of a licenser is necessary.

2.1 Downward Entailing

The first analyis, formulated by (Ladusaw 1979), claims that in order to license an NPI a sentence must be *Downward Entailing*. The notion of *Downward Entailing* can be explained as follows:

(11) *S1*: Every student must read this article.

(12) *S2*: Every first year student must read this article.

S1 entails *S2* as first year students belong to the set of all students, so if every student must read the article then first year students must read it too. This idea is formally defined in (13).

(13) *Ladusaw's generalization*
α is a trigger for negative polarity items in its scope iff α is downward entailing (where *trigger* is the expression that licensed ℵ).

This approach appeared to be fruitful and inspired a number of further research in the area ((Wouden T. 1994) and many other).

2.2 Nonveridicality

In (Giannakidou 2002) argues that the notion of Downward Entailing is not enough as there is a number of other contexts which are not Downward Entailing but still can license NPIs, such are pronouns like *hardly/barely* or non-monotone quantifiers like *almost nobody* or *nobody but* (**??**).

(14) a. Nobody but John saw anything.
b. Almost nobody saw anything.

(15) John hardly/barely talked to anybody.
a. John barely studied linguistics $\not\rightarrow$ John barely studied syntax.

Therefore, a notion of *(non)veridicality* was suggested:

(16) **Definition** - *(Non)veridicality for propositional operators*

a. A propositional operator *F* is veridical iff *Fp* entails *p*: $Fp \rightarrow p$; otherwise *F* is nonveridical.

b. A nonveridical operator *F* is *anti*veridical iff *Fp* entails *not p*: $Fp \rightarrow \neg p$.

According to (16), that a sentence is veridical if its proposition must be true for the sentence to be grammatical. If it is not, then it is nonveridicality or antiveridicality (which is basically direct negation).

(17) Yesterday Paul saw a snake. $\rightarrow$ Paul saw a snake.

(18) a. ?Did Paul see a snake? $\not\rightarrow$ Paul saw a snake.

b. Paul may have seen a snake $\not\rightarrow$ Paul saw a snake.

(19) Paul didn't see a snake. $\rightarrow$ It is not the case that Paul saw a snake.

To sum it up, regardless of what theoretical concepts are behind all these notions, a lot of researchers use nonveridical contexts as a base for research in NPI. All contexts we use in the current word are either nonveridical or antiveridical.

2.3 Classes of NPI

As we mentioned beforem it is important that every NPI has its own distributional pattern which means that not all NPIs are necessarily licensed in all the contexts mentioned above. If an NPI can occur only in the scope of direct negation and
or indirect negation. (Haspelmath 1997) proposed a map of contexts in which a polarized pronoun (or any other word, as appeared later) was occur.

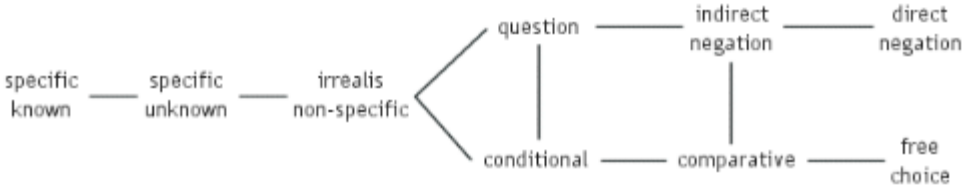


Figure 1: (Haspelmath 1997) map of licensing contexts

The point is that polarity is a scale and every item can be polarized to some extend. For this reason, three classes of NPIs by degree of polarization are usually

defined: strong, weak and everything in the middle of this scale. Strong NPIs are normally licensed by very small number of contexts on the right side of this scale (direct negation & indirect negation) while weak NPIs can be licensed in context like specific unknown.

3 Approaches to NPI extraction

While it seems more or less apparent that it is possible to use the tools of computational linguistics in order to extract polarity sensitive items, not so much research has been done in this area. However, computational approached showed great results when applied to German data.

3.1 Russian Polarity Sensitive Items

In (Апресян 2017) the results of semantic analysis of Russian polarized verbs and idioms with the use of corpus methods were presented. In this research a set of verbs and idioms extracted from dictionaries was tested with the use of Russian National Corpus[1] on occurrence in licensing contexts. The items included in the set had examples with negation in the dictionary articles. This criterion is based on the assumption that if a dictionary article contains an example of a word under negation then such a context is likely to be a typical usage for the item in question. Even though some of these items did not have any significant polarization but for most of them the approach appeared to be justified. As the result, a list of polarity sensitive verbs and idioms of different degrees of polarization was formed. However, most data was gathered manually which is not convenient if one wants to get larger lists of polarity items. Unfortunately, no computational methods to detect NPIs has ever been applied to Russian yet.

3.2 Hoeksema's Corpus Study of NPIs

One of the earliest attempts to use corpora in extracting NPIs and shows thhe potential of this approach was (Hoeksema 1997). Even though this article is not devoted to any computational tools, the results of using corpora in the field still appear to be rather fruitful .The article also mentions some potential problems with automatic extraction that may arise since not all the contexts can be precisely defined (20).

(20) a. You say anything, and I'll kill you.
b. *You said anything, and I killed you.

In (20) we have two sentences, one of the can license [anything] and the other cannot. It may seem that the tense of the verb is responsible for it but this

[1] **\url {http://www.ruscorpora.ru/}**.

is example is far more complicated. What really licenses *anything* in 20a is not the tense but the possibility to read the sentence as conditional while 20b does not provide such opportunity and both clauses can be regarded as veridical.

3.3 Collecting NPIs in German

Computational approach has been well tested on German data. In series of works by Frank Richter, Jan-Philipp Soehn, Timm Lichte and others, various computational tools were applied to corpora of German languages. In (Richter F. 2010) the main goal is to extend the *Collection of Distributionally Idiosyncratic Items* (CoDII), the list of all German NPI mentioned in literature which also includes search results for (Lichte 2005)

In (Soehn J.-P. 2010) an example of successful use of computational tools to extract NPIs is presented. The first step was to convert a corpus. The corpus in use was *Tübingen Partially Parsed Corpus of Written German* which consists of about 200 million words, though they used only a section of it which consisted of 5.8 million sentences from the years 1990 to 1998/

In order to obtain a list of NPI candidates, they used an association measure they called *context ratio* (CR) (3.3) which was computed for every relevant lemma *l* using its overall frequency *N* and the frequency $Nlic$ of configurations where *l* is in the scope of a licensing marker:

(21) $CR := \frac{Nlic}{N}$

They managed not only collect a large list of one-word NPIs but even a list of 'lemma chains' some of which in the end were defined as negative polarity expressions.

4 Method

In order to detect polarity sensitive we had to create a script that would perform the following tasks:

1. detecting licensing markers
2. defining which words are in its scope and which are not
3. collecting information about each candidate to count final quotient

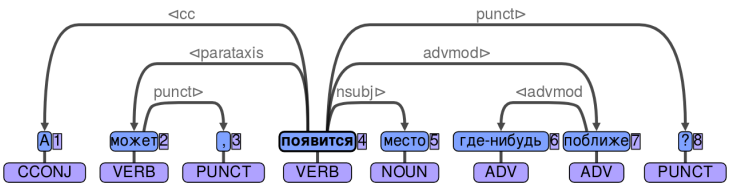
The most important constituents of our work are described below.

4.1 Corpora

The proper choice of corpora is of great importance for the current research, as every article in Section 3 mentions. Firstly, we needed corpora with a proper syntactic analysis as our items are supposed to be not just 'close' or 'in the same clause' with a licenser but in its scope, and for many languages, especially those of more or less free word order like Russian, it would be rather complicated to solve the problem without syntacticly annotated corpora. Fortunately, Universal Dependencies (UD) treebanks provide such annotation for over 67 languages. The annotation scheme uses universal inventory of functional categories which allows consistent annotation of constructions across languages. The UD project has an open community and by now there are treebanks for more than 60 languages and upcoming treebanks for 12 more languages. Thus, using UD format for this research not only allows us to use syntactic information for spotting NPIs but it would also simplify the process of adopting our tool for other languages, as long as treebanks for them exist, and allow us to conduct cross-linguistic research with the same tool, unlike Soehn J.-P. (2010) and (Richter F. 2010) and (Апресян 2017)

UD advantages and disadvantages

The main disadvantage of UD Corpora and Treebanks for our task was the fact that UD syntactic trees differ from standard binary syntactic trees in many ways.



А может, появится место где-нибудь поближе?
'Maybe, a place somewhere closer will appear?'

Figure 2: An example of a tree from Universal Dependencies treebank

The features which caused most of the limitations, are listed below and partially illustrated by Figure 2.

- **Non-binary.** Since the UD trees are not binary in most cases going down the tree may result in collecting items we do not need unless the relation between a head and a dependent are specified. However, in most contexts it is not always possible to predict these relations for each particular sentence.

- **Punctuation as nodes.** Punctuation marks are parts of the syntactic tree. It means that we can use them to find the borders of a clause or just ignore them as can be done in a case of 2 but dealing with cases like Figure **??** would require more thought especially if we want any code to be able to parse it.

- **Different types of dependencies.** Even though most relations are marked with a separate tag, these relations are not always comparable with common syntax. For example, an issue we would refer to as the address node problem in direct speech interrogative sentences. Though in common speech people address an interlocutor quite often, however, a syntactic representation of this wide-spread feature is still under debate(e.g. (Akkus F. 2018)). In our case it would be reasonable just to avoid 'address nodes' as they do not occure in scope of any relevant marker since they refer to the interlocutor of discourse who must be definite enough and, therefore, veridical. But the UD have a different view of it. Figure **??** shows that the question mark node is the child of the word in the 'address node' which is the child of the verb in the main clause (for a more detailed representation you can check Table 8 in Section 8). The problem here is that if we go up from the question mark node we will not reach for the rest of the interrogative clause and thus will not include it in the scope of question. For such particular cases we consider that every word between the question mark and dash are in the scope of question, and since for now we do take into account only verbs, adverbs and adjectives, this approach works fine but it must be modified if we want to be able to spot NPIs of other word classes.

UD Corpora for Russian

For the goals of the current research, we used corpora with approved UD annotation available on their github repositories (train, test and dev sets).

- **SynTagRus dependency treebank**[2] has more than 1 million tokens with news, nonfiction and fiction dated since 1960.

- **Taiga Corpus**[3]

  - Approved annotation[4]

[2] http://universaldependencies.org/treebanks/rusyntagrus/index.html

[3] https://tatianashavrina.github.io/taigasite

[4] https://github.com/UniversalDependencies/UDRussian-Taiga/tree/dev

  - Subset of not-approved annotation for posts and comments extracted from social media, particularly from VK social network[5] Since this part of Taiga annotation has no license the annotation for it is less trusted.

- **GSD**[6] consists of annotated acticles from Russian Wikipedia.

The script we wrote was applied to the corpora twice, with and without Taiga social media part. All in all, in the first case annotation for 1.5 million tokens (109351 sentences) was used, and adding annotation for text extracted from VK social network increased this number to more that 10 million tokens.

4.2 Licensing markers

For the goals of the current research we had to choose our own set of licensing contexts - and therefore, the markers of them.

Even though all the lists and scales of licensing contexts used for theoretical works are undoubtedly good as method, we could not use all of them in our work. The problem is that some of these licensers cannot be identified automatically with sufficient precision. That is why we had to limit the set of licensers to search in favor of better accuracy and precision. For this reason, applying Haspelmath's map approach does not seem to be possible in our case since we have no data for contexts or 'nodes' on his map like comparatives and free choice.

In the table 4.2 the types of contexts in and their markers are listed, illustrated with examples from Russian.

Type of context	**Marker**	**Example**
Direct Negation (DN)	negative particles and verbs	*Кроме села Веселого, ничего нет, - тактично не обращая внимания на её тон, ответил начальник.* -There is nothing except the Vesyoloje village, - answered the chef ingoring her tone.
Indirect Negation (IN)	negative conjunctions, negation in preceding clause, negative pronouns *некого, нечего*	*Я не хочу тебя терять.* -I don't want to lose you.

[5] `https://vk.com`

[6] `http://universaldependencies.org/treebanks/rugsd/index.html`

Question (QUEST)	question mark at the end of the sentence	*-А может, появится место где-нибудь поближе?* (Figure 2)
Conditional (COND)	*if*-sentences with *если*	*Если ты уйдёшь, я умру без тебя.*If you ever leave me I'll die without you.
Unreal Conditional (CONDIRR)	*если бы*	*Если бы я была свободна, я бы пришла.* -If I were free I would've come.
Unreal Conditional formed by (IRR) *бы* only	*бы*	*Я бы пошла, но буду занята.* - I'd like to go but I'll be busy.
Irrealis: Imperative (IMP)	Verbs in imperative form	*Спой что-нибудь.*
Irrealis: Modal contexts (MODAL)	Modal verbs	He must/can/should buy something
Restrictive particle *только* (RESTR)	*только*	

Each context required individual analysis of its structure. Unfortunately, due to the ambiguity of several UD relations, we had to narrow our research to polarity sensitive verbs.

4.3 Extraction

For every step described below the programming language in use was Python 3 including several open-source packages for it, mainly *conllu*[7] and *pandas*[8].

Preprocesing

The first step in the extraction process was to gather all annotated sentences and convert them to a Pandas Dataframe. The package *conllu* provides and opportunity to create dictionaries and trees but what for our script we needed to be able to quickly call every item in a sentence to be called by its index.

Detecting a licenser

We had a list of markers of licensers, so for each sentence we checked whether it has a licenser lemma or not. For most context we would just search for a

[7] `https://github.com/EmilStenstrom/conllu`
[8] `https://pandas.pydata.org/`

particular word, for instance, verbs like *мочь 'to can'*, *хотеть 'to want'*, *надо 'to have to'* would constitute modality in the sentence. If a sentences had a question mark at the end of it, we considered that every word in this sentence is in the scope of 'question.

Collection words in scope

If a sentence had any marker, we would check if there are any words of relevant word classes in its scope. Every such word would be added to a list or, if it is already in the list, the number of occurrences in a particular context would raise by one.

As the result, we got more that 19 thousand candidates. Obviously, not all of them are NPIs, so we applied CR from (Soehn J.-P. 2010) to them if the total number of occurrences is more than 10. Finally, we applied our version of (Апресян 2017)'s method to these candidates and got our list of polarity sensitive items.

5 Results

For the words *обинуясь*, *напастись*, *наздравствоваться*, *сроду*, *плошать*, *скупиться*, *взвидеть*, *притронуться*, *наготовиться*, *накупиться*, *дозваться*, *видаться*, *лежаться*, *писаться*, *навоевать*, *вытерпеть*, *уколупнуть*, *укупить*, *положить* and *постыдить* no occurences in relevant contexts were detected. For some of these words almost no occurrences were found at all since they words are rather rare. Their IPM[9] ratio is small even in the Russian National Corpora, which consists of 283431966 tokens.

$$ipm \text{ обинуясь} = \frac{162}{283431966} * 1000000 = 0.5716$$

$$ipm \text{ плошать} = \frac{121}{283431966} * 1000000 = 0.4269$$

$$ipm \text{ взвидеть} = \frac{117}{283431966} * 1000000 = 0.4128$$

$$ipm \text{ лежаться} = \frac{53}{283431966} * 1000000 = 0.1869$$

The words like *видаться 'see each other'*, even though their IPM is not that small, were not detected since their frequency has decreased by the time most texts of our UD Corpora were created.

Finally, the adverb *сроду 'ever'* was not detected since we searched not for any word in sentences with negation but only for those which are in the scope

[9] **items per million**.

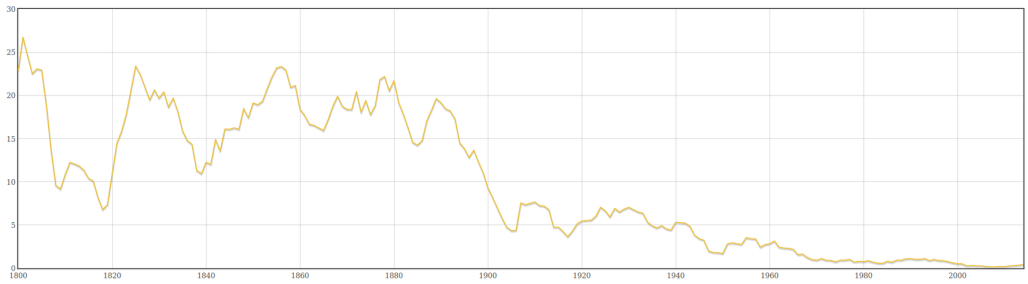


Figure 3: The frequency of *видаться* by year, according to the Russian National Corpus

of negation while in the case of *сроду* it is usually negation which is in the scope of *сроду*.

(22) У нас такого на всём заводе **сроду** не было.[Александр Солженицын. В круге первом, т.1, гл. 26-51 (1968) // «Новый Мир», 1990]
'Nothing like that has ever happened at our factory'

Nevertheless, we got enough occurences for 50 other items from Apresyan's list. For the subset of strong NPIs from this list, the CR 3.3 from (Soehn J.-P. 2010) was counted.

Word	**Occurrences in licensing contexts**	**All occurences in corpora**	**Context ratio**
житься	5	33	0.15151515151515152
запомнить	8	46	0.17391304347826086
обобраться	4	4	1.0
терпеться	8	8	1.0
подумать	309	1039	0.29740134744947067
посмотреть	3	9	0.3333333333333333
задуматься	21	97	0.21649484536082475
замедлить	4	10	0.4
преминуть	13	13	1.0
заладиться	2	3	0.6666666666666666
миновать	18	71	0.2535211267605634
надивиться	2	2	1.0
выносить	131	463	0.28293736501079914
клеиться	8	8	1.0
удосужиться	11	11	1.0

стерпеть	1	4	0.25
сидеться	3	4	0.75
ведать	51	108	0.4722222222222222
переваривать	22	23	0.9565217391304348
рыпаться	13	20	0.65

Table 2: Strong NPIs according to (Апресян 2017)

Some of this items are actually polarized in only one of their meanings or in a particular construction (5) (*подумать 'to think'*, *посмотреть 'to look'*, *ведать 'to be aware of'*, *миновать 'to avoid'*, *запомнить 'to remember'*, *задуматься 'to get lost in thought'*), that's why their CR is rather low.

(23) a. Но наш герой миновал шестнадцатый этаж, пятнадцатый, десятый, восьмой... [Очередной подвиг отечественного воровства (2003) // «Криминальная хроника», 2003.06.10]
'But our character passed the sixteenth floor, the fifteenth, the tenth, the eigth...'

b. — Чему быть, того **не** миновать, — сказала Ольга и как-то ушла в себя. [Вячеслав Пьецух. Шкаф (1997)]
*'-What is meant to happen, can**not** be avoided, - said Olga and got closed and indifferent.'*

However, for those items which are mostly or only used in a sense with polarization, the CR is at least 0.5. For this reason, we decided to check all our candidates which have a CR equal or bigger that 0.8. Our script does not detect contexts like *future* or pragmatic licensers which means that numerous occurrences in apparently licensing contexts are neglected. However, we consider that if a word occurs many times at least in the contexts we were able to detect often enough then it should have some significant degree of polarization, even though several occurrences are missed. As the result, we got the following list of items:

Since our script does not recognize different meanings, every item of appropriate CR was checked manually in the Russian National Corpus in a ways similar to (Апресян 2017). As the result, we managed to define a number of other polarity sensitive items, some of which are presented below (others can be found on my github page[10].

[10] https://github.com/mbibaeva/NPI_extraction

Word	**Translation**	**Our CR**	**RNC with neg**	**Degree of polarity**
моргнуть	to blink	$\frac{12}{12} = 1$	58%	middle
перечить	to contradict	$\frac{8}{8} = 1$	85%	strong
клеиться2	to go well (lit. to glue)	$\frac{8}{8} = 1$	98.5%	strong
смешить hline покладать	to make laugh to put	$\frac{81}{87} = 0.93$ $\frac{32}{33} = 0.97$	86% 100%	strong strong
утруждать	to bother	$\frac{21}{24} = 0.87$	99%	strong
афишировать	to advertise, to show off	$\frac{41}{46} = 0.89$	93%	strong
гнушаться	to abhor	$\frac{35}{37} = 0.95$	87.5%	strong
вздумать	to take in one's head	$\frac{15}{15} = 1$	74.5%	middle
тешиться	to have fun	$\frac{6}{6} = 1$	61%	middle
отвлечься	to get distracted	$\frac{6}{6} = 1$	50.5%	middle
гнаться	to chase	$\frac{19}{20} = 0.95$	57.7%	middle
полениться	to be too lazy to do smth	$\frac{51}{55} = 0.93$	77.3%	middle
обделываться	to fail	$\frac{32}{34} = 0.94$	50%	middle
выклевать	to peck out	$\frac{15}{16} = 0.94$	51.8%	middle
высовываться	to stick out	$\frac{18}{20} = 0.9$	50%	middle
побояться	to be too afraid	$\frac{34}{41} = 0.83$	54%	middle

All in all, we managed to get around 50 new items of different degrees of polarization. That may seem good, however, that shows that the accuracy of existing script cannot be trusted as these items were selected from more than 400 hundreds of verbs with CR > 0.5 and more than 10 occurrences.

6 Conclusion

Automatic extraction of polarity sensitive items is not an impossible task, as can be proven by (Richter F. 2010), (Soehn J.-P. 2010) and this current work. However, it order to not just collect candidates but to make some accurate analysis for the more strict rules would be needed as well as advances measures of statistic

analysis.

7 Bibliography

References

Akkus F., Hill V. (2018). “The Speaker in Inverse Vocative”. In: *Proceedings of the 35yth West Coast Conference on Formal Linguistics*. Ed. by Bennet Wm.G. Ed. by Lindsay Hracs. Ed. by Denis Ryan Storoshenko, pp. 49–58.

Giannakidou, A. (2002). “Licensing and Sensitivity in Polarity Items: from Downward Entailment to (Non)veridicality”. In: *Chicago Linguistic Society* 38-2. Ed. by Mary Andronis, pp. 29–54.

Haspelmath, M. (1997). *Indefinite Pronouns*. Oxford Studies in Typology and Linguistic Theory. Oxford University Press.

Hoeksema, J. (1997). “Corpus study of negative polarity items”. In: *Jornades de corpus linguistics* IV-V, pp. 931–952. URL: `http://odur.let.rug.nl/~hoeksema/docs/barcelona.html`.

Ladusaw, W.A. (1979). “Negative Polarity as Inherent Scope.” Ph.D. thesis. University of Texas, Austin.

Lichte, T. (2005). “Corpus-based acquisition of complex negative polarity items”. In: *Proceedings of the Tenth ESSLLI Student Session*. Ed. by Gervain J.

Richter F. Fritzinger F., Weller M. (2010). “Who Can See the Forest for the Trees? Extracting Multiword Negative Polarity Items from Dependency-Parsed Text”. In: *JLCL*, pp. 83–110.

Soehn J.-P. Lichte T., Trawinski B. (2010). “Spotting, collecting and documenting negative polarity items”. In: 28, pp. 931–952.

Wouden T., van der (1994). “Negative contexts”. Ph.D. thesis. University of Groningen.

Апресян, В.Ю. (2017). “Отрицательная и положительная поляризация: семантические источники”. Russian. In: *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной конференции ”Диалог”(Москва, 31 мая - 3 июня 2017г.)* 16(23). Ed. by Селегей В.П..

8 Appendix

Word	Percentage	Class
Житься	100	strong
Запомнить2	100	strong
Обобраться	100	strong

Терпеться	100	strong
Подумать	100	strong
Посмотреть	100	strong
Задуматься	99	strong
Замедлить	99	strong
Обинуясь	99	strong
Преминуть	99	strong
Заладиться	96	strong
Миновать	95	strong
Пара	95	strong
Напастись1	94	strong
Надивиться	92	strong
Наздравствоваться	92	strong
Сроду	91	strong
Выносить	90	strong
Плошать	90	strong
Скупиться.	90	strong
Клеиться	89	strong
Взвидеть	89	strong
Удосужиться	88	strong
Стерпеть	87	strong
Притронуться	86	strong
Сидеться	86	strong
Наготовиться	86	strong
Ведать	82	strong
Переваривать	82	strong
Накупиться	82	strong
Рыпаться	81	strong

Table 4: Strong NPIs from (Апресян 2017)

1	-	-	PUNCT	_	_	2:punct
2	Как	как	ADV	_	Degree=Pos	0:root
3	с	с	ADP	_	_	5:case

4	моим	мой	DET	_	Case=Ins Gender=Neut Number=Sing	5:det
5	делом	дело	NOUN	_	Animacy=Inan Case=Ins Gender=Neut Number=Sing	2:obl
6	,	,	PUNCT	_	_	7:punct
7	товарищ	товарищ	NOUN	_	Animacy=Anim Case=Nom Gender=Masc Number=Sing	2:parataxis
8	началь-ник	началь-ник	NOUN	_	Animacy=Anim Case=Nom Gender=Masc Number=Sing	7:appos
9	?	?	PUNCT	_	_	8:punct
10	-	-	PUNCT	_	_	7:punct
11	спросила	спросить	VERB	_	Aspect=Perf Gender=Fem Mood=Ind Number=Sing Tense=Past VerbForm=Fin Voice=Act	2:parataxis
12	она	она	PRON	_	Case=Nom Gender=Fem Number=Sing Person=3	11:nsubj
13	,	,	PUNCT	_	_	16:punct
14	сознатель-но	сознатель-но	ADV	_	Degree=Pos	16:advmod
15	не	не	PART	_	_	16:advmod
16	называя	называть	VERB	_	Aspect=Imp Tense=Pres VerbForm=Conv Voice=Act	11:advcl

17	Семена	семен	PROPN	_	Animacy=Anim Case=Acc Gender=Masc Number=Sing	16:obj
18	Еремее- вича	еремее- вич	PROPN	_	Animacy=Anim Case=Acc Gender=Masc Number=Sing	17:flat
19	по	по	ADP	_	_	20:case
20	имени	имя	NOUN	_	Animacy=Inan Case=Dat Gender=Neut Number=Sing	16:obl
21	-	-	PUNCT	_	_	22:punct
22	отчеству	отчество	NOUN	_	Animacy=Inan Case=Dat Gender=Neut Number=Sing	20:nmod
23	.	.	PUNCT	_	_	22:punct

Table 5: Example of UD sentence analysis