Réseaux de neurones pour une détection automatique des NPI

Ekaterina Kolos[1,3] Pascal Amsili[2,3]
(1) Université Paris Nanterre, 200 Av. de la République, 92000 Nanterre, France
(2) Sorbonne-Nouvelle, 8 avenue de Saint-Mandé, 75012 Paris, France
(3) Laboratoire Lattice, 1 rue Maurice Arnoux, F-92120 Montrouge, France
ekaterina.kolos@sorbonne-nouvelle.fr, pascal.amsili@ens.fr

RÉSUMÉ

Cet article présente un travail en cours sur la détection des NPI pour les textes anglais et présente les résultats obtenus dans la première partie de ce projet. Nous présentons à la fois un corpus annoté, une version préliminaire d'un système d'étiquetage et ses premiers résultats.

ABSTRACT

Automatic Negative Polarity Item Detection

This paper introduces a work in progress on NPI Detection for English texts and presents the results obtained in the first part of this project. We present in this paper an annotated corpus, a preliminary version of a tagging system, and its first results.

MOTS-CLÉS : polarité, monotonie, NPI, FCI, étiquetage de séquences, classification multiclasse.

KEYWORDS: polarity, monotonicity, NPI, FCI, sequence tagging, multiclass classification.

1 Context and motivation

Negative Polarity Items (NPI) are lexical units like the English *any* that are only grammatical in a limited number of contexts, also called licensing contexts. The most common licensing context is negation (1), which explains why they are called NPI, but many other contexts have been shown to licence NPIs, like interrogative sentences, the restrictor of a universal quantifier, or the antecedent clause of a conditional sentence, etc. (Homer, 2020).

(1) (a) *John has *any* friend(s).
(b) John does not have *any* friends.

The list of NPIs (lexical units or constructions) is very heterogeneous, both syntactically (pronouns, determiners, adverbs, NPs...) and semantically. Some NPIs belong to closed classes, which means that we can make a complete list of them (determiners/pronouns...), but other NPIs are built around nouns denoting a small quantity (*a clue*, *a finger*...), so that we have an open list of NPIs. In addition, licensing contexts are also very diverse, and semanticists are still working to establish the list of licensing environments (see Appendix for the list we adopt here), and to try to determine what those environments have in common.

We are concerned in this work with the automatic detection of NPIs and identification of their licensing contexts. Detecting NPIs is not a current task identified as such in the natural language processing community, however, we consider that it would be interesting to have a way to detect NPIs with a good quality, for several reasons. It would allow linguists to collect data for theoretical investigations on polarity (and also free choice items, see below), as well as monotonicity and negation. It might also prove useful to get additional features for the NLP task of detecting negation and its scope. Downstream applications like natural language generation or text correction may also benefit from a proper identification of polarity items and contexts. Furthermore, NPI detection may be used as a diagnostic classifier (Hupkes *et al.*, 2018) when assessing models' capabilities to learn complex semantic and syntactic concepts (Jumelet & Hupkes, 2018; Jumelet *et al.*, 2021; Bylinina & Tikhonov, 2021).

The task of detecting NPIs is made more complex by the fact that several lexical units that have a NPI function (like *any* or *ever*) can also be used in contexts where they get a different interpretation, dubbed "free choice" (2) (Fauconnier, 1978; Giannakidou, 1998).

(2) You can take *any* book you like.

A separate class of polarity items has been proposed, the so-called free-choice items (FCI), that for some researchers constitute a separate group (Fauconnier, 1978) and for others are a subclass of NPI (Homer, 2020). These items can be licensed, among others, by imperatives, modal verbs, comparatives and superlatives. This class comprises items that behave only as FCI, like *whatever*, but also items which can be used both as FCI and as NPI.

The task we are proposing thus requires not only that NPIs are identified as well as their licensing contexts, but also that they are distinguished from FCI.

Identifying FCI and NPI is not that easy as it might seem at first glance. First of all, these items can be ambiguous with non polar expressions, like (3), where *on earth* is not used as an NPI, *vs.* (4), where the expression is a typical minimizer used as an NPI.

(3) The Arctic is experiencing some of the most rapid and severe climate change *on earth*.

(4) Why *on earth* did you leave me ?

Second, multiple items of different classes can be present in one phrase, e.g. (5), where the first *any* is a weak NPI and the second one an FCI.

(5) Is there *any* specific food I might find in *any* pet shop ?

Finally, minimizers, that can be found alone (4), together with an FCI (6) or an NPI (7), constitute an open class of lexical items.

(6) I would appreciate *any* insight on this *at all*.

(7) They don't know *anything* about wine *at all*.

All these items have a particular set of licensing environments in common, although some of them have more limited distribution than others. We believe it might be of interest to train a system capable to identify those licensing environments in order to :

1. disambiguate known NPI when they are present in a sentence ;
2. identify new NPI that haven't been seen yet ;
3. explore to what extent neural models can grasp complex syntactic and semantic dependencies.

We model the task as a sequence labelling task, with a BIO-scheme. Instead of having only two classes (NPI *vs.* non-NPI) we decided to make a further distinction among NPIs : we separate FCI and NPI, minimizers from the rest of NPIs, and we make a distinction between weak and strong NPIs.

As for the licensing context, we chose to encode the type of context in the tag the sequence receives.

The tags in BIO-scheme follow the pattern `B/I-NPI_Type-Licensor_Type`. This results in 128 possible different B/I-tags, 40 of which were present in the train data.

In the rest of this paper, we present the dataset that we created (§ 2) and we give the first results that were obtained with BiLSTM and BERT (§3). We close the paper with a discussion and perspectives.

2 Annotated dataset

The first stage of our work was to produce an annotated dataset that we could use to train and evaluate our models. We started a pilot annotation campaign, with two annotators, and a rather rich tagset since we not only distinguished FCI from NPI, but also subclasses within NPIs. Sixteen licensing environment classes were distinguished (see Appendix).

We pre-selected a number of NPI candidates – a list that is by no means exhaustive but that could serve as a starting point (see Appendix).

English texts from Universal Dependencies (Nivre *et al.*, 2016) were annotated, making possible further use of syntactic information. A total of 1596 sentences [1] out of a total of 38068 sentences were found to contain one or more of the manually pre-selected NPI candidates and added to the dataset (see Appendix for a detailed table). In case two NPI candidates were present in the sentence, two separate datapoints were created, which resulted in 1734 separate datapoints. [2]

The small percentage of sentences with NPI we found here falls in line with the numbers demonstrated in Jumelet & Hupkes's study (Jumelet & Hupkes, 2018) where a total of 301.836 (2.69%) sentences containing any form of *any* (*anybody, anyone, anymore, anything, anytime*, and *anywhere*) were extracted from 11.213.916 sentences of their Google Books corpus.

We made the following decisions when annotating our data :

1. Minimizers are NPI with the following properties : they cannot be licensed by non-monotonous contexts (c.f. *Exactly two students did anything* vs **Exactly two students lifted a finger to help*) ; they are grammatical in affirmative sentences when negating a negation (c.f. *A : You don't give a damn about my problems. B : But I do give a damn !*) ; they can be licensed by some modal verbs in the sense of irrealis (*You could've lifted a finger to help* vs **You could've done anything to help.*) (Sailer, 2021) ; following Homer, we also add *at all* to this group (Homer, 2020).
2. *Any* is an NPI in negative sentences, when it disappears with the change of polarity. Thus, 8 is an NPI and 9 is an FCI.

[1] 2 were removed during the annotation phase

[2] In the abstract initially submitted we reported having worked with 1885 data samples ; we eventually reduced the number of items we process during this first stage, eliminating, e.g., *quite* used as NPI in examples like *He's not quite sure about it* != ¬*He's quite sure about it.*

(8) Mary isn't trying *anything* to get Mark back. = ¬ Mary is trying something to get Mark back.

(9) Mary isn't ready to try [just] *anything* to get Mark back. = ¬ Mary is ready to try just anything to get Mark back.

3. *any* is, similarly, an NPI in other negative licensing environments, such as negation in the main clause, implicit negation.
4. *any* is, similarly, an NPI in questions and indirect questions.
5. *any* is, similarly, an NPI in antecedents of conditionals.
6. *any* is, similarly, an NPI when licensed by a restrictor of a universal quantifier.
7. *any* is an NPI when licensed by *only*.
8. *any* and *ever* are FCI when licensed by superlatives and comparatives, as well as *too*-phrases. Having this in mind, we also consider FCI *yet* in *the best I've seen* ***yet***.
9. *any* is an FCI when licensed by imperatives and indirect imperatives.
10. *any* and *ever* are FCI when licensed by restrictors such as relative clauses, which define the set of objects from which one can 'freely choose'.
11. *below*, *before*, *prior to* are considered a separate licensing environment and license NPI and not FCI *any* : *Before he could do* ***anything****, the car crashed into the tree.*
12. the negated *any-... but* is considered an NPI : *I haven't seen* ***anything*** *but care and consideration.*
13. *any* in idiomatic *if any*, *if X is* ***any*** *guide* is annotated as NPI : *John has very few friends, if* ***any*** and is considered to be licensed by the *antecedent of a conditional* environment.
14. *any*- items in *any-... of* are annotated as FCI : *We can't rely on* ***any*** *of them.*
15. *anything*, *whatever* in idiomatic *or* ***anything***, *or* ***whatever*** are annotated as FCI.

We also understand that what we annotate as FCI is not homogeneous. A canonical example of freedom of choice would be, e.g., (10). We also annotated as FCI, however, (11), which literally means '*Every* piece of information will be appreciated' and (12), where *any* is a part of an idiom.

(10) Put in a heater and set it to *anywhere* between 78-82.

(11) *Any* and all information will be appreciated.

(12) She was not having *any* of that.

We consider a more linguistically informed classification, e.g. taking into account semantic properties such as downward-entailment (Ladusaw, 1979), non-/antiveridicality (Giannakidou, 1998; Zwarts, 1998), or Strawson entailment (Von Fintel, 1999) a direction for future work.

3 Experiments

To explore how NPIs can be extracted based on the data we annotated, we preprocess the datapoints, merging, where necessary, multiple tags for one sentence, and learn two different models on the resulting BIO-scheme.

We explore two subtasks : first, the system has to be able to predict whether specific tokens of a sentence form an item of interest for us or not (i.e. identify the NPI's boundaries and disambiguate). Second, we try to predict the item's class : [FCI - weak - strong - minimizer] x 16 licensing environments.

For both purposes, we use two models : one is a **BiLSTM** - a simple architecture of random embeddings followed by a bidirectional LSTM network and a linear layer with a softmax to predict the most probable class for each token. We expect this model to learn licensing environments to the left (*I wonder if anyone has any suggestions*) and to the right (*Anyone have any suggestions ?*) of the NPI candidate.[3] The second model we try is **BERT**, based on a pretrained BERT model from huggingface (`bert-base-uncased`) with a classification unit on top. Our data consists of texts of different genres and domains, and we use the most general BERT model without any domain-specific fine-tuning of the embeddings prior to training the classifier.[4]

For the first subtask, the BiLSTM model correctly predicts NPI boundaries (i.e. the entire sequence of 'B', 'I', 'O' tags) in 88.1% of test sentences, the BERT model - in 96.9% cases. A baseline tagger relying solely on a list of NPIs and matching substrings from it showed 68.1% accuracy on the same data. This metric is later referred to as *acc str* in subtask 2.

Table 1 shows the results for subtask 2 : multiclass classification. To better evaluate the performance of our multiclass classifier we compute the following metrics :

1. *acc str* : estimates the number of entirely correctly predicted sentences ; this is quite strict, since an error in one tag corresponds to an incorrect prediction for the whole sentence, although the sentence might have multiple correctly predicted NPI candidates ;
2. *acc tag* : estimates the number of correctly predicted tokens in the whole test dataset, without taking sentences into account : so, if the test dataset contained 2 sentences each of 3 tokens, these would constitute 6 separate datapoints, each for every token ;
3. *acc bi* : same as *acc tag*, but now only for non-'O' tags, i.e. the 'B' and 'I' tags, where the model had to predict the class of the NPI and its licensor ; we need this metric because the *acc tag* metric is biased due to the large proportion of 'O' tags which are easier to predict ; this third metric only evaluates how many NPI and licensor classes were predicted correctly ;
4. weighted average Precision, Recall and F-Score are also provided.

The quantitative estimation shows clearly that the BERT model outperforms the LSTM pipeline.

model	*acc str*	*acc tag*	*acc bi*	P	R	F1
BiLSTM	0.656	0.982	0.572	0.63	0.62	0.62
`bert-base-uncased`	0.831	0.991	0.787	0.85	0.86	0.83

TABLE 1 – subtask 2 : Multiclass classification results

Apart from the quantitative estimation above, we tried to qualitatively estimate our models by asking them to tag new examples inserted manually. We were particularly interested to know if the models could tag minimizers that they had not previously seen. For example, the minimizer *a hoot*, or minimizers based on swear words never occurred in our training data. In our BERT experiments we

[3] A combination of SGD optimizer, cross-entropy loss function, and dropout gave the best results. Bidirectional LSTM proved more capable of learning different licensing contexts.

[4] The results we list below were obtained with Adam optimizer, learning rate of 1e-05, 10 epochs.

could identify such previously unseen NPI, as in (13) ; in other experimental settings BERT only tagged *a* as a *B*-tag (beginning of a minimizer). In any case, the model did not tag *a hoot* in (b) where it is not used as a minimizer, neither did the models consider (c) an example of a minimizer, although it is of similar syntactic structure.

(13) unseen NPI : positive (a) and negative (b, c) examples :

(a)	John	does	not	give	**a**	**hoot**
	O	O	O	O	**B-NPI**	**I-NPI**
(b)	The	owl	gave	a	loud	hoot
	O	O	O	O	O	O
(c)	John	does	not	have	a	cat
	O	O	O	O	O	O

The licensing contexts that were better learnt (first of all, because they were better represented) were negation, direct and indirect questions, comparatives, and antecedents of conditionals (a detailed classification report can be found in the Appendix).

4 Conclusion and Future Work

In this work, we introduced a new annotated dataset for NPI and FCI categorization, as well as a first attempt to categorize these items based on our annotation with the help of deep learning tools.

The models we build seem to be capable of grasping syntactic and semantic information on NPI without any explicit syntactic hints.

The models used here leave room for improvement, for example by adding a CRF layer for consistent *B* and *I* tags, by using semantically aware embeddings or combining BERT and LSTM. Better quality might be achieved by learning the licensor type and the NPI type independently, e.g. with a two-head BERT model.

A further direction of future work would be to annotate a bigger dataset with balanced classes, as well as formalize the annotation guide, invite more annotators and estimate the inter-annotator agreement. The current system could be used to select potential data for this new corpus. One could also explore capabilities of multilingual models, like multilingual BERT, in transferring knowledge of NPI licensing from one language to another, which could prove useful in low-resource scenarios.

Acknowledgments

This work was accomplished during the first author's master internship at the Lattice lab. It was supported by the labex EFL (Empirical Foundations of Linguistics, ANR-10-LABX0083). We would like to thank our intern Thi Hien Linh Nguyen for participating in the first annotation pass, as well as Ilya Kuryanov for assisting in proof-reading and editing the manuscript. We thank the anonymous reviewers for their comments on the first version of this short paper.

Références

BYLININA L. & TIKHONOV A. (2021). Transformers in the loop : Polarity in neural models of language. *arXiv preprint arXiv :2109.03926*.

FAUCONNIER G. (1978). Implication reversal in a natural language. In *Formal semantics and pragmatics for natural languages*, p. 289–301. Springer.

GIANNAKIDOU A. (1998). *Polarity sensitivity as (non) veridical dependency*, volume 23. John Benjamins Publishing.

HOMER V. (2020). *Negative Polarity*, In *The Wiley Blackwell Companion to Semantics*, p. 1–39. John Wiley & Sons, Ltd. DOI : https ://doi.org/10.1002/9781118788516.sem057.

HUPKES D., VELDHOEN S. & ZUIDEMA W. (2018). Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, **61**, 907–926.

JUMELET J., DENIĆ M., SZYMANIK J., HUPKES D. & STEINERT-THRELKELD S. (2021). Language models use monotonicity to assess npi licensing. *arXiv preprint arXiv :2105.13818*.

JUMELET J. & HUPKES D. (2018). Do language models understand anything ? on the ability of lstms to understand negative polarity items. *arXiv preprint arXiv :1808.10627*.

LADUSAW W. A. (1979). *Polarity Sensitivity as Inherent Scope Relations.* The University of Texas at Austin.

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIC J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N. *et al.* (2016). Universal dependencies v1 : A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 1659–1666.

SAILER M. (2021). Minimizer negative polarity items in non-negative contexts.

VON FINTEL K. (1999). Npi licensing, strawson entailment, and context dependency. *Journal of semantics*, **16**(2), 97–148.

ZWARTS F. (1998). Three types of polarity. In *Plurality and quantification*, p. 177–238. Springer.

Appendix

NPI licensing environments distinguished in this work

1. `1_negation` : negation (negative particles, conjunctions, prepositions (*without*, *against*, *unless*) : *John left home without eating **any** breakfast*, *John won't leave unless he finds **anything** useful*) ;
2. `2_hidden negation` : hidden negation, i.e. non-affirmative verbs (*I doubt that John ate **any** breakfast*), negative predicates (*unlikely* : *John is unlikely to eat **any** breakfast*), other expressions which we informally identify as having some negative meaning, e.g. *It was a big to-do to find **anyone** who knew it = One couldn't easily find **anyone** who knew it* ;
3. `3_quantifier of small quantity` : negative quantifiers, or quantifiers of small quantity (*few/little* : *Few commuters **ever** take the train to work*, *Little can be done to change **anything** for the better*) ;
4. `4_exactly` : non-monotonous quantifier licensing weak NPI (not found in our data, category reserved for further annotation) : *Exactly two students had **any** success with this task.*
5. `5_question` : questions (*Has **anyone** already figured out the answer ?*) ;
6. `6_indirect question` : indirect questions (*I wonder if **anyone** already figured out the answer* ; *I don't want to comment on whether they did any of that.*) and subjunctives : (*John is sorry that Bill said **anything** against Paul*) ;
7. `7_antecedent of conditional` : antecedents of conditionals (*If **anyone** notices **anything** unusual, it should be reported*) ;
8. `8_restrictor of universal quantifier` : restrictors of universal quantifiers (*Every customer who had **ever** purchased anything in the store was contacted*) ;
9. `9_comparatives and superlatives, too-phrases` : comparatives and superlatives, *too*-phrases, *first*, *last* (*John is taller than **any** other employee*), *John is too short to see **anything***) ;
10. `10_imperative` : imperatives (*Take **any** book you like*) ;
11. `11_indirect imperative` : indirect imperatives (*I want you to take **any** book you like*) ;
12. `12_relative_clause_or_other_restrictor` : relative clauses and other restrictors (*John talked to **any** woman who came up to him*) ;
13. `13_modal_irrealis` : (*John could have lifted **a finger** to help, but he didn't*) ; not present in the data, reserved for further annotation ;
14. `16_temporal` : a category added later for licensing through *before*, *after*, *prior to* etc (*Before he could do **anything**, the car hit the tree*) ;
15. `14_other_free_choice` : basically all free-choice usages except those explained by imperatives and explicit restrictors (***Anyone** can do it*) ;
16. `15_other_NPI` : a category for all other contexts licensing NPI proper, for example, *only* : *Only John brought **any** friends.*

Manually pre-selected NPI Candidates

item	NPI	FCI	ambiguous	was annotated
any	NPI : weak	yes	no	yes
any way	NPI : weak	yes	no	yes
anybody	NPI : weak	yes	no	yes
anyone (any one)	NPI : weak	yes	no	yes
anyhow	NPI : weak	yes	no	yes
anything	NPI : weak	yes	no	yes
anywhere	NPI : weak	yes	no	yes
ever	NPI : weak	yes	yes	yes
either	NPI : strong	no	yes	yes
yet	NPI : strong	yes*	yes	yes
a bean	NPI : minimizer	no	yes	yes
a bit	NPI : minimizer	no	yes	yes
a bite	NPI : minimizer	no	yes	yes
a clue	NPI : minimizer	no	yes	yes
a damn	NPI : minimizer	no	no	yes
a drop	NPI : minimizer	no	yes	yes
a finger	NPI : minimizer	no	yes	yes
a fly	NPI : minimizer	no	yes	yes
a note	NPI : minimizer	no	yes	yes
a penny	NPI : minimizer	no	yes	yes
a single word	NPI : minimizer	no	yes	yes
a thing	NPI : minimizer	no	yes	yes
a word	NPI : minimizer	no	yes	yes
all that	NPI : minimizer	no	yes	yes
an eye	NPI : minimizer	no	yes	yes
an inch	NPI : minimizer	no	yes	yes
at all	NPI : minimizer	no	yes	yes
whatsoever	NPI : minimizer	no	yes	yes
whatever	no	yes	yes	yes
whenever	no	yes	yes	no
wherever	no	yes	yes	no
whoever	no	yes	yes	no
whichever	no	yes	yes	no

TABLE 2 – Our NPI Inventory

Number of sentences with NPI from our list in UD corpora

Corpus	NPI Candidates	Sentences	From Total Sentences
en_ewt-ud-dev.conllu	77	72	2001
en_ewt-ud-train.conllu	699	632	12543
en_ewt-ud-test.conllu	90	86	2077
en_gum-ud-dev.conllu	36	33	843
en_gum-ud-train.conllu	231	216	5660
en_gum-ud-test.conllu	27	26	894
en_atis-ud-dev.conllu	16	16	572
en_atis-ud-train.conllu	99	99	4274
en_atis-ud-test.conllu	12	12	586
en_lines-ud-dev.conllu	65	62	1032
en_lines-ud-train.conllu	169	158	3176
en_lines-ud-test.conllu	62	57	1035
en_partut-ud-dev.conllu	3	3	156
en_partut-ud-train.conllu	103	88	1781
en_partut-ud-test.conllu	13	13	153
en_pronouns-ud-test.conllu	0	0	285
en_pud-ud-test.conllu	25	23	1000

TABLE 3 – Number of sentences with NPI extracted from the Universal Dependencies English Corpora

Classification reports

	LSTM Results				BERT Results			
	P	R	F1	support	P	R	F1	support
NPI_FCI_10	0.00	0.00	0.00	4	0.67	0.50	0.57	4
NPI_FCI_11	0.00	0.00	0.00	2	0.00	0.00	0.00	2
NPI_FCI_12	0.00	0.00	0.00	6	1.00	0.17	0.29	6
NPI_FCI_14	0.54	0.48	0.51	27	0.81	0.96	0.88	27
NPI_FCI_9	0.93	0.78	0.85	18	1.00	1.00	1.00	18
NPI_minimizer_1	0.44	0.50	0.47	8	0.44	0.50	0.47	8
NPI_minimizer_5	0.00	0.00	0.00	1	0.00	0.00	0.00	1
NPI_minimizer_7	0.00	0.00	0.00	1	0.00	0.00	0.00	1
NPI_strong_1	1.00	0.83	0.91	6	1.00	0.83	0.91	6
NPI_strong_5	0.00	0.00	0.00	0	0.00	0.00	0.00	0
NPI_weak_1	0.80	0.77	0.79	31	0.80	0.77	0.79	31
NPI_weak_2	0.33	0.33	0.33	3	0.50	0.33	0.40	3
NPI_weak_3	0.00	0.00	0.00	1	0.00	0.00	0.00	1
NPI_weak_5	0.62	0.84	0.71	19	0.90	1.00	0.95	19
NPI_weak_6	0.33	1.00	0.50	1	1.00	1.00	1.00	1
NPI_weak_7	0.80	0.73	0.76	11	0.91	0.91	0.91	11
micro avg	0.62	0.62	0.62	139	0.88	0.86	0.87	139
macro avg	0.36	0.39	0.36	139	0.63	0.56	0.57	139
weighted avg	0.63	0.62	0.62	139	0.85	0.86	0.83	139

TABLE 4 – Test results of the LSTM model (on the left) the BERT model (on the right)