HAL
open science

From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers

Ingrid Falk, Delphine Bernhard, Christophe Gérard

To cite this version:

Ingrid Falk, Delphine Bernhard, Christophe Gérard. From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. LREC - The 9th edition of the Language Resources and Evaluation Conference, May 2014, Reykjavik, Iceland. hal-00959079

HAL Id: hal-00959079

https://inria.hal.science/hal-00959079v1

Submitted on 2 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers

Ingrid Falk, Delphine Bernhard, Christophe Gérard

LiLPa - Linguistique, Langues, Parole
EA 1339, Université de Strasbourg
{ifalk,dbernhard,christophegerard}@unistra.fr

Abstract

In this paper we present a statistical machine learning approach to formal neologism detection going some way beyond the use of exclusion lists. We explore the impact of three groups of features: form related, morpho-lexical and thematic features. The latter type of features has not yet been used in this kind of application and represents a way to access the semantic context of new words. The results suggest that form related features are helpful at the overall classification task, while morpho-lexical and thematic features better single out true neologisms.

Keywords: neologisms, semi-automatic neologism detection, classification, French, topic modeling

1. Introduction

The work we present in this paper is concerned with the semi-automatic detection of French neologisms in online newspaper articles.

The system we develop, called *Logoscope*, retrieves newspaper articles from several RSS feeds in French on a daily basis. Using exclusion lists it identifies unknown words, which are then presented to a linguist to decide which of these are valid neologisms. For example, Table 1 shows the most frequent unknown words, collected on July 12, 2013 and resulting from this procedure. The table also illustrates a major drawback of this method. Clearly most of these forms are not interesting neologism candidates: in many cases they are not even valid words and a linguist expert would have to tediously scan a large part of the list before finding interesting candidates.

lmd (18)	twitter/widgets (7)	india-mahdavi (3)
pic(this (18)	garde-à (6)	kilomètresc (2)
lazy-retina (9)	ex-PPR (4)	geniculatus (2)
onload (9)	pro-Morsi (4)	margin-bottom (2)
onerror (9)	tuparkan (4)	politique» (2)
amp;euro (7)	candiudature (3)	...

Table 1: The most frequent unknown words collected on 2013-07-12. Word frequency is shown in parentheses.

In this study we investigate methods to select among the detected unknown forms those which most probably represent interesting neologism candidates.

We address this task by casting it into a supervised classification problem. The classification is based on different types of features extracted from the collected newspaper articles. A further objective of this paper is to explore which kinds of features are most helpful at detecting the most interesting neologism candidates.

The paper is organized as follows. We first briefly present previous work and relate our study to these earlier efforts. We then detail the resources and methods our experiments are based on and finally present their results.

2. Previous Work

Methods for the automatic or semi-automatic identification of neologisms mainly target the coinage of new words or changes in part-of-speech (*grammatical neologisms*), while semantic neologisms are only seldom dealt with.

2.1. Identification of Grammatical Neologisms

Grammatical neologisms correspond to attested word forms which are used with a new part-of-speech. This is a rather rare phenomenon, which may nevertheless cause problems for POS taggers. Janssen (2012) describes NeoTag, a language independent tagger and lemmatizer, which takes grammatical neologisms explicitly into account. After tagging, the tagged corpus is compared with a lexicographic resource, in order to detect words whose categories are different in the corpus and in the resource.

Given that grammatical neologisms are scarce, we do not address this phenomenon for the time being, but rather focus on the detection of new word forms.

2.2. Detection of New Word Forms

For this task, two different types of methods may be distinguished:

- Methods based on lists containing known words in the target language, which are used to identify unknown words. These lists are usually called *exclusion lists* ;
- Methods relying on various statistical measures or machine learning applied to diachronic corpora.

The use of exclusion lists is by far the most common method. For French, the POMPAMO tool (Ollinger and Valette, 2010) uses an exclusion list made of the French lexicon MORPHALOU 2.0 (Romary et al., 2004), a list of named entities and lexicons provided by the user. In addition to the lexicons, the tool uses filters which detect non-alphanumeric characters, numbers and composed word forms. Issac (2011) also describes several filters, aimed at eliminating unwanted neologism candidates not found in

the exclusion list. The first filter eliminates non words by looking for bigrams and trigrams of characters which are not found in French. The second filter targets words which are concatenated due to missing spaces and looks for all the possible combinations in a dictionary. Finally, spelling errors are identified by finding corrections with the Levenshtein distance.

Systems relying on exclusion lists have also been developed for languages other than French. CENIT – Corpus-based English Neologism Identifier Tool – (Roche and Bowker, 1999) uses additional filters which aim at detecting proper nouns. For German, the Wortwarte platform[1] collects neologisms on a daily basis (Lemnitzer, 2012).

All these methods rely on simple heuristics and require that the candidates be manually validated by an expert.

Statistical measures can be employed when neologisms are studied from a diachronic point of view. Garcia-Fernandez et al. (2011) use the Google Books Ngrams to identify the date of coinage of new words, based on cumulative frequencies for a time period going from 1700 to 2008. The cumulative frequency curve of neologisms is exponential. New words are identified when this cumulative frequency exceeds a given threshold. Cabré and Nazar (2011) also exploit this property to identify neologisms in Spanish newspaper texts.

Finally, machine learning can be employed to automatically classify neologism candidates. Stenetorp (2010) presents a method to rank words extracted from a temporally annotated corpus. The features correspond to the number of occurrences of the word in the corpus, the distribution of its occurrences over time, the points in time when the word is observed, lexical cues (quotes, enclosing in specific HTML tags), the presence in dictionaries, and spell-checking.

The main limitation of diachronic methods is that they necessitate a corpus covering a large time span. Moreover, neologisms can only be detected some time after their first occurrence.

The work we present here is an attempt to combine the two main approaches. We use exclusion lists as filters but also apply machine learning techniques to select the most probable neologism candidates. Since we expect the users of our system to be interested in the creation of new words for different reasons and to have different views on the phenomenon, we did not rely on a very specific definition of neologisms. Moreover, our goal is to detect neologisms on their first occurrence, on a daily basis. As a consequence, we do not include diachronic features in our study.

2.3. Detection of Novel Word Senses

Here, the aim is to detect new word senses for attested word forms. It is only recently that computational – automatic or semi-automatic– approaches have been proposed for this task. One reason may be its difficulty: in many cases, new word senses may be rather infrequent and thus challenging to detect with corpus-based methods (Cook and Hirst, 2011; Lau et al., 2012).

Several different strategies may be employed: (i) study compound word forms, (ii) rely on local lexico-syntactic contexts, or (iii) analyse the wider themes of the text.

[1] http://www.wortwarte.de

For German, Lemnitzer (2012) proposes analyzing novel compound word forms to detect new senses of the base word forms. However, this idea has not been automatized yet.

Cook and Hirst (2011) compare lexico-syntactic representations of word contexts in corpora corresponding to two time periods. They then use synthetic examples built from the Senseval dataset to evaluate their system for the identification of semantic change.

Contextual information is also used by Boussidan and Ploux (2011) who propose several clues for the detection of semantic change, relying on a co-occurrence based geometrical model. Their model is able to represents thematic clusters but the approach is not fully automatic. Following the same idea, Lau et al. (2012) apply topic modeling in order to perform word sense induction, and thus detect novel word senses. In this case, topic modeling is not applied to whole documents but to contexts representing target words. For the identification of novel word senses, two corpora are used: one reference corpus, to represent standard uses and known word senses, and a newer corpus with unknown word senses. Words senses are induced on both corpora and then words are tagged with their most likely sense. A novelty score is then computed, based on the senses observed in the reference corpus and in the newer corpus.

In our work, we do not focus on the detection of novel word senses, as we only target the identification of new word forms. Our work is nevertheless related to previous work in this domain in that we use topic modeling in order to represent the thematic context of unknown words forms, under the assumption that some thematic contexts are more prone to the coinage of words than others.

3. Experimental Setting

3.1. The Data

For our experiments we collected a corpus of French RSS feeds, from the newspapers and on the dates shown in Table 2.

Total number of articles: 2,723	
Newspapers: Le Monde (659), Libération (504), l'Équipe (594), Les Echos (956)	
Dates: July 12, 16, 19, 23, 25, 30, August 1 2013	
Total number of forms (tokens):	51,000
Unknown forms (types):	692
Valid forms (types):	265
Invalid forms (types):	427
True neologisms (types):	81

Table 2: Newspaper corpus collected for our experiments (in parentheses: the number of articles collected for each newspaper).

We preprocessed the corpus and only kept the journalistic content. The articles were then segmented in sentences and tokenized using the TinyCC tools.[2] Finally the tokens were filtered using an exclusion list based mainly on Morphalou (Romary et al., 2004) and Wortschatz (Biemann et

[2] http://wortschatz.uni-leipzig.de/~cbiemann/software/TinyCC2.html

al., 2004). We also detect Named Entities using a slightly modified version of the CasEN NER utility (Maurel et al., 2011). Finally, we end up with a list of 692 unknown words after filtering. The list of unknown words was further manually categorized into the following classes:

- Plausible words whose form is correct in French. Non words mainly correspond to spelling errors (e.g. *candiudature*), foreign words (e.g. *lazy-retina*) and source code which was not stripped from the HTML file (e.g. *onload*).
- Neologisms which are the candidates our system should eventually extract. Plausible words which are not neologisms correspond mainly to words which should be included in the exclusion lists.[3]

We performed two series of (supervised) classification experiments. For both the items to be classified are the unknown words not found in the exclusion list (cf. Figure 1). In one set of experiments the positive examples used to train the classifier are the plausible words (the negative examples being the remaining unknown words). This setting is called *Plausible words* in the following.
In the second set of experiments (called *Neologisms*) the positive examples used in training are the validated neologisms and the negative examples the remaining unknown words.

3.2. Features

We explored the effect on the classification of three groups of features: *form* features, *morpho-lexical* features and *thematic* features, an overview of which is presented in Table 3. The **form features** are the most obvious features to be used in such a classification task. They are related to the form or construction of the string at hand, and are language independent. Table 3a shows some examples of these features.
Table 3b gives an overview of the main **morpho-lexical features**. First, these features check whether particular prefixes and suffixes are present and whether some characters indicate particular languages.[4] Second, we assess the probability that the form might be a spelling error by using the aspell tool.[5] This tool suggests a list of known spellings close to the tested form. The feature we use is the Levenshtein distance to the best guess (or 100 if there is no suggestion).
Based on the observation that unknown forms often arise from missing white space we use a further group of morpho-lexical features to check whether other known forms are possibly included in the form at hand. This group of features is derived from the results of the Aho-Corasick string matching algorithm (Aho and Corasick, 1975)[6] which suggests a list of known forms present in the unknown form at hand. Obviously these features depend on morpho-lexical characteristics of French.
Since one of the goals of the *Logoscope* project is to provide means for observing the creation of new words in an enlarged textual context we also focus on the influence of **thematic features** on the automatic identification of neologisms (Table 3c). Our hypothesis is that these features supply interesting additional information not provided by form related and morpho-lexical features. Besides the obvious *Newspaper* feature, we attempt to capture the thematic context using topic modeling (Steyvers and Griffiths, 2007). Topic models are based on the idea that documents are mixtures of topics, where a topic is a probability distribution over words. Given a corpus of documents, standard statistical techniques are used to invert this process and infer the topics (in terms of lists of words) that were responsible for generating this particular collection of documents. The learned topic model can then be applied to an unseen document and we can thus estimate the thematic content of this document in terms of the inferred topics.
In our experimental setting we use topic modeling as follows. We first assemble a set of *general journalistic themes* from a large collection of newspaper articles. Based on these topics we then estimate the thematic content of the larger textual context of the unknown words we investigate. Several studies ((Blei and Lafferty, 2009; Hoffman et al., 2010; Hoffman et al., 2013)) show that in general tens or hundreds of thousand documents are needed for a thorough thematic analysis of this kind and that the number of extracted topics is of 100-300. However, for the preliminary study presented here we collected 4,755 articles from the newspapers shown in Table 3c and restricted the number of extracted topics to 10.
Each unknown form is associated with 20 topic features (2 features for each of the extracted 10 topics). The first set of 10 features is obtained by first concatenating the sentences containing the unknown form. We then apply the obtained topic model to this text and obtain topic proportions estimating its thematic content. The unknown form is then associated with these topic proportions, representing the weight of each topic in its phrasal contexts. To compute the values of the second set of the 10 topic features we apply our topic model to each article containing the unknown form. For each topic, the unknown form is then associated with the maximal proportion for this topic found in the articles containing the form. Thus, these features represent the predominant topics of the articles containing the unknown forms.

3.3. Methodology

The most straightforward way to cast our neologism detection problem into a supervised classification task is to consider the 81 validated neologisms as the positive examples and the remaining 611 unknown words as the negative

[3] However, since it is not always clear whether a word is a neologism, arguably valid unknown words are worth keeping and observing.

[4] We used the `Lingua::Identify` perl script to this end: `http://search.cpan.org/~ambs/Lingua-Identify-0.56/lib/Lingua/Identify.pm`.

[5] `http://aspell.net/`

[6] We used the Perl implementation available at the url `http://search.cpan.org/~vbar/Algorithm-AhoCorasick-0.03/lib/Algorithm/AhoCorasick.pm`

Length: Number of characters
Whether the form contains particular signs, digits, whether it is capitalised,
Relative and absolute frequency wrt. to documents and sentences

(a) Examples of form related features.

Language
Whether the form contains characters indicating a particular language (French, English, German or Spanish, 5 features).
Prefix/suffix
0 or 1 depending on whether a particular prefix or suffix is present. Prefixes: ultra-, super-, dé-, ré-, ... (69 in total) Suffixes: -iste, -ation, -isme, -itude, ... (30 in total)
Spelling
spell-checker (aspell): Is it a spelling error? (2 features) Aho-Corasick algorithm (Aho and Corasick, 1975): Does the form contain other known forms? (4 features)

(b) Morpho-Lexical features.

Topics	
10 topics extracted from 4,755 newspaper articles: Le Monde (898), Libération (269), La Libre (1,784), Presse Europe (690), Le Journal du dimanche (206), Rue89 (212), Slate (74), L'Équipe (892)	
Topic features: 10 features, proportion of topic in document	
Documents	Feature value
articles containing form	max. topic proportion
concatenation of sentences containing form	topic proportion
Newspaper: The newspaper(s) the form appeared in.	

(c) Thematic features.

Table 3: Features

examples in the training data. This is one of the classification scenarios (called *Neologisms*) we investigate using the three groups of features presented in Section 3.2. We saw two reasons to also experiment with a second classification scenario. First, as mentioned earlier, it is not always clear when exactly an unknown word turns into a neologism. Second, the strongly imbalanced data suggests that the resulting classification model may possibly be inaccurate. In this second scenario (called *Plausible words*) we consider the 265 plausible words as positive training examples (and the remaining 427 as negative examples) and explore the impact of the different feature sets on the resulting classifications. The two classification scenarios and the involved data items are illustrated in Figure 1.

3.4. Classifier

We used an SVM classifier as implemented in LibSVM (Chang and Lin, 2011) and the Weka tookit (Hall et al., 2009).

In the *Neologisms* classification setting we accounted for the strongly imbalanced data by oversampling the positive class.[7] We then used the LibSVM classifier implemented in Weka with the default settings.[8] In the *Plausible words* classification scenario the class distribution is less imbalanced, so we used the resampling technique proposed in Weka. We also used a radial basis kernel and performed a grid search for the optimal cost and γ parameters.[9]

We perform a 10-fold cross-validation and report precision, recall and F-measure for each class separately and for both classes taken together. We also report the number of detected true neologisms.

When applying the *Neologisms* classification technique, the detected true neologisms are a direct result of the classification process. In contrast, the *Plausible words* classification scenario results in those unknown forms which are considered to be plausible words, but which are not necessarily true neologisms. Since ultimately we are interested in these, we also rank the found positive items using the probability estimates output by the LibSVM tool. We then evaluate how many of the true neologisms were among the 81[10] words with highest probability.

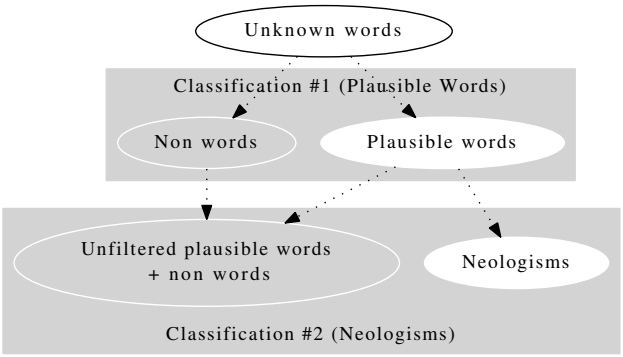

Figure 1: Classification scenarios

4. Results and Discussion

Results. Table 6 shows the results obtained in the two classification scenarios: *Plausible words* on the left hand side of Table 6 and *Neologisms* on the right hand side. We performed the classifications with combinations of the groups of features described in Section 3.2.: form related features, morpho-lexical features and thematic features. The results are given in terms of precision, recall and F-measure for each class and for the overall classification obtained through 10-fold cross validation. We also present the correctly identified neologisms: in the *Neologisms* scenario these are simply the items classified correctly as positive, whereas in the case of the *Plausible words* setting these are

[7] Oversampling is a classification technique which helps to deal with imbalanced data. The instances of the minority class are duplicated in order to obtain approximately as many instances as in the majority class. We used the Weka cost sensitive classifiers to achieve this.

[8] An exponential kernel with cost 1 and $\gamma = 0$.

[9] We used the scripts provided with the LibSVM tool (Chang and Lin, 2011) for this.

[10] Recall that there are 81 validated neologisms to be detected.

agroécologiste (0.961798)	Etat-départements (0.921507)
anti-alcoolisme (0.959942)	nationalistes-révolutionnaires (0.904493)
anti-salazariste (0.953199)	démission-suprise (0.891366)
non-audition (0.939236)	auto-obscurcissant (0.87521)
multiactivité (0.92963)	constructeur-carrossier (0.870512)
restaurant-snack-bar (0.925892)	ultra-présent (0.868557)

Table 4: Top entries in the reordered unknown word list obtained in the *Plausible word* setting with the *lex* feature set, the configuration which resulted in the largest number of true neologisms. In parentheses the probability that the unknown form is a plausible word.

the validated neologisms ranking in one of the top 81 positions.

First the results show that using the machine learning techniques presented here, the unknown words filtered by our system can be reordered and presented to a human expert in a more meaningful way. Table 4 shows a possible reordering produced by our system.

All classifiers show a reasonable performance with respect to the measure which is most relevant for our application: the recall for the positive class, which reflects the number of identified true neologisms.[11]

Regarding the overall F-measure, the behaviour of both sets of classifiers was similar. The best results were achieved with attribute sets involving the form features (form and thematic features for the *Plausible words* classification and form attributes only for the *Neologisms* classification). For both scenarios the classifications with overall best F-measure identified the less validated neologisms.

The settings which provide the best balance between global F-measure and the number of identified neologisms are *form+lex* for *Plausible words* and *form+lex+theme* for *Neologisms*.

The results suggest that the features used in our experiments were sufficiently powerful to support the *Neologism* classification scenario outweighing the unbalanced data and the difficulty of the classification task.

Groups of features. The morpho-lexical features proved to be helpful for neologism detection in both classification scenarios. Interestingly, while in the *Plausible words* setting form features also had a strong impact, in the *Neologisms* scenario it was the thematic features which played a more important role. First this confirms our intuition that the thematic context is more helpful at the detection of new words than at the detection of plausible words: plausible words may appear in any context whereas this outcome suggests that thematic context is to some extent linked with word creation. This finding is in line with an important line of work in (Textual) Linguistics where word creation is found to correlate with certain discourse types and textual genres (see for example (Cabré et al., 2003; Elsen, 2004; Elsen and Dzikowicz, 2005; Ollinger and Valette, 2010)). Second, this highlights the benefit of using topic modeling for a more compact representation of thematic content insofar as this representation gives some access to the semantic context of unknown forms going beyond the more limited co-occurrence windows. This aspect, as shown in Section 2. is rarely taken into account, theoretically or practically, by recent neologism detection utilities.

Qualitative discussion. For a qualitative analysis of the effect of the various types of features on the selection of valid new words we applied the classification models based on the *form*, *lex*, *theme* and *form-lex-theme* feature groups on our data and examined the number of correctly identified neologisms and for each group the five best scoring unknown words and the five best scoring validated neologisms. Table 5 shows the results of these experiments. First we observe that the *form* features help identify particularly long neologisms, and those containing a hyphen (first line). The second line shows that the words scoring best with the *lex* features are mainly compositions (with or without hyphen) or contain a prefix (*anti, non*). With respect to the contextual features (*theme*) we observe on the positive side that they permit the detection of neologisms with no prominent property (*agnélise, retricoté*), but on the negative side these thematic features (*theme*) seem to favour the selection of words which are not plausible considering traditional formation rules for French word forms. A closer look at the neologisms detected through the *theme* features but not via *lex* and *form* features confirmed their ability to select neologisms with less characteristic forms. Some neologisms identified by the *theme* classifier, but not by the *lex* and *form* classifier are: *accrobranches, caricatureurs, conflicté, frenemies, instinctivores* . . . .

5. Conclusion and Future Work

In this paper we presented a supervised statistical machine learning method which helps at the identification of new words in online newspaper publications. First we identified words in online newspaper articles which were not present in an exclusion list. Using support vector machines these unknown words were then ranked according to the probability of their being valid neologism candidates, based on different features extracted from the newspaper articles. Despite the relatively small amount of data, our experiments showed that this way the unknown words could be re-arranged in a more meaningful way (than randomly or by frequencies) thus facilitating an analysis by a linguistic expert or lexicographer.

We examined the impact on the classification outcome of three groups of features: *form* features, related to the construction of the word and language-independent, *morpho-lexical* features based on lexical characteristics and therefore language dependent and finally *theme* features which are meant to reflect the larger textual context of the unknown word and have not yet been used to this end (to our knowledge).

The best F-measure for the global classification task (results for positive and negative classes combined) was achieved based on the *form* related features in conjunction with the *theme* features, showing the relevance of this type of attributes for neologism detection. However, the configura-

[11] While the overall F-measure in the *Plausible words* classification scenario is acceptable, it is lower in the case of the *Neologism* classification scenario. These results could possibly be improved by also learning the kernel parameters for the oversampled training data. However the absolute F-measure is less important in our setting since we are more interested in the impact of the attribute sets.

Features	#neos	top 5 valid neologisms	top 5 (neologism?)
form	37	supermédiateur, doublevédoublevé-doublevé, auto-diagnostiqués, néo-célibataires, sur-monétisation	styliste-couturière (no), E-DÉTOURNEMENTS (yes), supermédiateur (yes), garde-à (no), doublevédoublevédoublevé (yes)
lex	48	agroécologiste, multiactivité, auto-obscurcissant, neo-retraité, macrostabilité	agroécologiste (yes), anti-alcoolisme (no), anti-salazariste (no), non-audition (no), multiactivité (yes)
theme	48	e-détournements, partenadversaires, hollandisme, retricoté, agnélise	tuitte (no), e-détournements (yes), schlopp (no), gesagt (no), schloppa (no)
form-lex-theme	60	pagano-satanisme, auto-diagnostiqués, neo-retraité, agroécologiste, e-détournements	ultra-présent (no), Etat-départements (no), anti-alcoolisme (no), pagano-satanisme (yes), watts-étalons (no)

Table 5: Top 5 predictions when applying the model.

tion with the best overall F-measure produced the less validated neologisms.[12] Most neologisms were identified using attribute sets involving lexical and thematic features. This highlights the importance of these features and suggests that the classification technique could be further improved by their better integration.

Since we found the "thematic" features to be of great importance for the neologism detection and documentation we plan to expand our work on detecting and documenting general purpose, journalistic themes, using the topic modeling techniques described in this paper. However, as mentioned in Section 3.2., for a thorough thematic analysis and documentation based on general journalistic themes we need to apply topic modeling on a much larger and more "representative" corpus of newspaper articles. In addition further investigation is needed to determine the influence of other parameters as for example the vocabulary – which words are most meaningful in our setting and therefore are most relevant in the topic selection process, or what number of topics to choose. Further, for documenting the unknown words an interpretation (or labeling) of the topics would be important.

In future work we also plan to exploit other "textual" features which have been found in linguistic studies to play an important role in word creation. Some of these features are:

- The position of the unknown word in the text (Loiseau, 2012).
- The journalistic genre – our hypothesis is that some journalistic genres are more favourable to word creations than others. For this however first a (systematic) categorisation of genres and second a method is needed to better identify the journalistic genre of a newspaper text.

A further interesting research direction is to explore more fine-grained measures in order to better assess the influence of the different features on the classification result. In (Lamirel et al., 2013) the authors propose versatile feature analysis and selection methods allowing to improve the classification results independent of the classification method and to accurately investigate the influence of each feature on the classification process.

[12] While the number of identified validated neologisms corresponds to the recall in the case of the *Neologism* experiments, this is not the case for the *Plausible words* experiments.

Acknowledgements

We thank Romain Potier-Ferry for his contributions. This work was financed by an IDEX 2012 grant from the Université de Strasbourg.

6. References

Aho, Alfred V. and Corasick, Margaret J. (1975). Efficient string matching: an aid to bibliographic search. *Commun. ACM*, 18(6):333–340, June.

Biemann, Christian, Bordag, Stefan, Heyer, Gerhard, Quasthoff, Uwe, and Wolff, Christian. (2004). Language-Independent Methods for Compiling Monolingual Lexical Data. In Goos, Gerhard, Hartmanis, Juris, Leeuwen, Jan, and Gelbukh, Alexander, editors, *Computational Linguistics and Intelligent Text Processing*, volume 2945, pages 217–228. Springer Berlin Heidelberg.

Blei, David M. and Lafferty, J. (2009). Topic models. *Text mining: classification, clustering, and applications*, 10:71.

Boussidan, Armelle and Ploux, Sabine. (2011). Using topic salience and connotational drifts to detect candidates to semantic change. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), Oxford*.

Cabré, Teresa and Nazar, Rogelio. (2011). Towards a new approach to the study of neology. In *Neology and Specialised Translation 4th Joint Seminar Organised by the CVC and Termisti*.

Cabré, M. T., Domènech, M., Estorpà, R., Freixa, J., and Solé, E. (2003). L'observatoire de néologie: conception, méthodologie, résultats et nouveaux travaux. *L'innovation lexicale*, pages 125–147.

Classification								
	Plausible words				Neologisms			
form, lex, theme								
class	**Prec**	**Rec**	**F**	**corr. neos**	**Prec**	**Rec**	**F**	**corr. neos**
pos	0.848	0.618	**0.715**		0.181	0.827	0.297	
neg	0.800	0.932	0.861		0.958	0.512	0.667	
both	0.818	0.813	0.806	24	0.868	0.548	0.625	67
form, lex								
class	**Prec**	**Rec**	**F**	**corr. neos**	**Prec**	**Rec**	**F**	**corr. neos**
pos	0.822	0.565	0.670		**0.192**	0.778	0.308	
neg	0.776	0.925	0.844		0.952	0.573	0.716	
both	0.794	0.788	0.778	30	0.864	0.597	0.669	63
form, theme								
class	**Prec**	**Rec**	**F**	**corr. neos**	**Prec**	**Rec**	**F**	**corr. neos**
pos	0.814	**0.687**	0.745		0.160	0.531	**0.346**	
neg	0.825	0.904	0.863		0.912	0.638	0.751	
both	0.821	0.822	**0.818**	22	0.826	0.625	0.693	43
form								
class	**Prec**	**Rec**	**F**	**corr. neos**	**Prec**	**Rec**	**F**	**corr. neos**
pos	0.635	0.405	0.494		0.190	0.481	0.273	
neg	0.702	0.857	0.772		0.915	0.733	0.814	
both	0.676	0.686	0.666	20	0.832	0.704	**0.752**	39
lex								
class	**Prec**	**Rec**	**F**	**corr. neos**	**Prec**	**Rec**	**F**	**corr. neos**
pos	0.777	0.466	0.582		0.132	0.827	0.227	
neg	0.737	0.918	0.818		0.927	0.288	0.440	
both	0.752	0.746	0.728	**34**	0.836	0.350	0.415	67
theme								
class	**Prec**	**Rec**	**F**	**corr. neos**	**Prec**	**Rec**	**F**	**corr. neos**
pos	**0.850**	0.389	0.534		0.129	**0.889**	0.225	
neg	0.719	0.958	0.822		0.938	0.217	0.353	
both	0.769	0.742	0.712	22	0.844	0.295	0.338	**72**
lex, theme								
class	**Prec**	**Rec**	**F**	**corr. neos**	**Prec**	**Rec**	**F**	**corr. neos**
pos	0.844	0.557	0.671		0.136	0.877	0.236	
neg	0.776	0.937	0.849		0.945	0.275	0.426	
both	0.802	0.793	0.781	29	0.851	0.345	0.404	71

Table 6: Classification results

Chang, Chih-Chung and Lin, Chih-Jen. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Cook, Paul and Hirst, Graeme. (2011). Automatic identification of words with novel but infrequent senses. In *PACLIC*, page 265–274.

Elsen, Hilke and Dzikowicz, Edyta. (2005). Neologismen in der Zeitungssprache. *Deutsch als Fremdsprache: Zeitschrift zur Theorie und Praxis des Deutschunterrichts für Ausländer*, (2):80–85.

Elsen, Hilke. (2004). *Neologismen: Formen und Funktionen neuer Wörter in verschiedenen Varietäten des Deutschen*. Gunter Narr Verlag.

Garcia-Fernandez, Anne, Ligozat, Anne-Laure, Dinarelli, Marco, and Bernhard, Delphine. (2011). Méthodes pour l'archéologie linguistique: datation par combinaison d'indices temporels. *Actes du septième DÉfi Fouille de Textes*.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, and Witten, Ian H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Hoffman, Matthew D., Blei, David M., and Bach, Francis R. (2010). Online Learning for Latent Dirichlet Allocation. In *NIPS*, pages 856–864.

Hoffman, Matthew D., Blei, David M., Wang, Chong, and Paisley, John. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Issac, Fabrice. (2011). Cybernéologisme : Quelques outils informatiques pour l'identification et le traitement des néologismes sur le web. *Langages*, 183:89–104.

Janssen, Maarten. (2012). NeoTag: a POS Tagger for Grammatical Neologism Detection. In *LREC*, page 2118–2124.

Lamirel, Jean-Charles, Cuxac, Pascal, Hajlaoui, Kafil, and Chivukula, Aneesh Sreevallabh. (2013). A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data. In *International Workshop on Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models (QIMIE)*, Australie, April.

Lau, Jey Han, Cook, Paul, McCarthy, Diana, Newman, David, and Baldwin, Timothy. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, page 591–601.

Lemnitzer, Lothar. (2012). Mots nouveaux et nouvelles significations — que nous apprennent les mots composés ? *Cahiers de lexicologie. Néologie sémantique et analyse de corpus*, 1(100):105–116.

Loiseau, Sylvain. (2012). Un observable pour décrire les changements sémantiques dans les traditions discursives : la tactique sémantique. *Cahiers de lexicologie*, 1(100):185–199.

Maurel, Denis, Friburger, Nathalie, Antoine, Jean-Yves, Eshkol, Iris, and Nouvel, Damien. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement Automatique des Langues*, 52(1):69–96.

Ollinger, Sandrine and Valette, Mathieu. (2010). La créativité lexicale : des pratiques sociales aux textes. In *Actes del I Congrés Internacional de Neologia de les llengües romaniques*, volume Publicacions de l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF), pages 965–876, Barcelona, Spain.

Roche, Sorcha and Bowker, Lynne. (1999). Cenit : Système de détection semi-automatique des néologismes. *Terminologies nouvelles*, (20):12–16.

Romary, Laurent, Salmon-Alt, Susanne, and Francopoulo, Gil. (2004). Standards going concrete: from LMF to Morphalou. In *Workshop Enhancing and Using Electronic Dictionaries*, Geneva, Switzerland.

Stenetorp, Pontus. (2010). Automated Extraction of Swedish Neologisms using a Temporally Annotated Corpus. Master's thesis, Royal Institute of Technology (KTH), Stockholm, Sweden, March.

Steyvers, Mark and Griffiths, Tom, (2007). *Probabilistic Topic Models*. Lawrence Erlbaum Associates.