The Logoscope: a Semi-Automatic Tool for Detecting and Documenting French New Words From the Linguistic Project to the Web Interface

Ingrid Falk, Delphine Bernhard, Christophe Gérard

The Logoscope: a Semi-Automatic Tool for Detecting and Documenting French New Words

From the Linguistic Project to the Web Interface

Ingrid Falk Delphine Bernhard Christophe Gérard

July 11, 2018

In this article we present the design and implementation of the *Logoscope*, the first tool especially developed to detect new words of the French language, to document them and allow a public access through a web interface. This semi-automatic tool collects new words daily by browsing the online versions of French well known newspapers such as Le Monde, Le Figaro, L'Équipe, Libération, La Croix, Les Échos. In contrast to other existing tools essentially dedicated to dictionary development, the *Logoscope* attempts to give a more complete account of the context in which the new words occur. In addition to the commonly given morpho-syntactic information it also provides information about the textual and discursive contexts of the word creation; in particular, it automatically determines the (journalistic) topics of the text containing the new word. In this article we first give a general overview of the developed tool. We then describe the approach taken, we discuss the linguistic background which guided our design decisions and present the computational methods we used to implement it.

Contents

1 Introduction: New Words, NLP and the Logoscope

Like other areas of linguistics, the study of new words has benefited from the development of natural language processing techniques in the last twenty years. In this research area, the digital revolution has significantly changed the way to collect the data necessary for empirical research (Mejri and Sablayrolles, 2011; Humbley and Sablayrolles, 2016): The traditional practice of collecting new words *by reading* texts (newspapers, literature, scientific and technical texts, etc.),[1] has been supplemented by an automatisation of the collection process. In fact, today there are many computer tools capable of searching through large amounts of texts (often newspapers) in order to automatically detect the newly created words in various languages (German, Catalan, Spanish, English, French, etc.).

What is a new word? is, however, one of the first questions to ask when addressing this topic. This is a very difficult question, for which, to our knowledge there is no clear consensual answer in the scientific community.

A naïve but plausible standpoint is to consider new words to be all the lexical units which are absent from all the lexicons that were developed in the history of a language. Besides being hardly feasible, a reason for which this approach is problematic is that arguably these newly created words are not really present or important in the contemporary "textual public use". Often they are hapaxes which were used in one text only (*eg.*, *quadrambulation*, *sucraphone*, etc.[2]) or in few texts (*eg.*, *bredictable*). In other cases their social diffusion is currently too low to be able to speak of collective *use* (*eg.*, *artwashing*). Other word creations, like *brexit*, clearly reflect a societal actuality, are widely publicised but therefore clearly not new. On the other hand some words which have already been integrated in a general dictionary (as *phubbing* and *hot take* which were included in this year's Oxford Dictionary) or in a specialised glossary (terminology) could arguably be considered as new by a typical newspaper reader.

From the standpoint of the *Logoscope* ideally both hapaxes and words like *brexit* should be monitored because their distribution in public texts over time provides insights about the life cycle of new words and about the interaction between texts in different domains of the public space and lexical innovation.

Computational natural language processing methods can easily collect from online texts those forms which are not in a predefined list of known words. However, when using this approach one would encounter several obstacles. Firstly, it is not feasible to compile all the known words of a language (from all dictionaries or terminologies) into such a predefined list. Secondly, online texts are noisy and computational methods are never perfect, so the extracted unknown forms would surely not always be valid words of the considered language. For these reasons we consider that a manual validation will always be necessary. But even a linguist expert can not possibly know all the valid words of a language and there is no generally accepted definition of what a new word is. Therefore ultimately the decision whether a word is new is necessarily subjective, the linguist expert has to rely on the context and on her knowledge and intuition.

Based on these considerations, our tool was designed to first identify unknown forms using a necessarily incomplete list of known words. Then a linguist expert decides which of these are interesting word creations, based on lexicographic criteria but also taking into consideration the societal and cultural context in which the form occured.

In this article we adopt a pragmatic position and talk of "new words", "recent words" or "recently / newly created words" rather than "neologism" and "neology", "lexical creation", "neonym", "constructed word", "lexical innovation", etc. Indeed, this raises well known terminology issues (Guilbert, 1975; Rey, 1976; Boulanger, 2010; Sablayrolles, 2006; Pruvost and Sablayrolles, 2012; Cabré, 2015, 2016) that we cannot discuss here, especially since this would require an extensive reflection on the meta-language of linguistics (Neveu, 2008; Swiggers, 2010), which is well beyond the scope of this work. However, for ease of use

[1] For the French language, the manual detection method has produced the *Base d'Observation et de Recherche des Néologismes*, `http://www.atilf.fr/borneo`, ATILF - CNRS & Nancy Université.

[2] Theses hapaxes were created by *newologism*(@NewThingFriends) on Twitter: "Quadrambulation: Walking around the block" (June 5, 2016), "Sucraphone: A speaker of sweet nothings" (July 7, 2016).

we will sometimes employ the word "neologism", but exclusively as a synonym to "new word".

These thousands of new words detected each year may be used for several purposes. They are essential for making general language dictionaries, for driving language policies (Quebec, Spain, etc.), for the observation of science and technology, or to measure the vitality of languages and dialects (Cabré, 2000). Furthermore, lexical creation constantly testifies, within a language area, of economic and political events, of the formation of social identities (slang, etc.), or changes in attitudes (*eg.*, (Borde, 2016)), and more broadly, cultural history (Calvet, 2016). Finally, most new words express a rhetorical dimension that is crucial in political communication, advertising or creative writing.

However, to use the collection of new words to this end it is necessary to dispose of a tool addressed not only to lexicographers and terminologists, but also to other user communities (journalism, translation, marketing, social sciences, etc.). For these users the appearance of new words may also be relevant since it correlates with events of interest for their particular community (technical, societal, political, sportive, etc.). This is precisely the particular ambition of the framework presented in this paper, the *Logoscope*, and what distinguishes it from other existing IT tools. To this end the new word creations are documented not only using well-known lexicographic categories describing the form itself, but also by taking into account their historical and co(n)textual environment. In the *Logoscope* framework we achieved this by first developing a principled semi-automatic method to set up a collection of higher-level, coarse-grained topics reflecting what journalistic articles are generally about. This way we could describe the newspaper articles in terms of the topics they were handling and in the same time assess the thematic context in which the word creation in a particular article occurred.

The article is organised as follows. In Section 2 we give a general overview of the *Logoscope* framework. Section 3 is an account of previous work on the automatic detection and documentation of new words, while Section 4 focuses on the linguistic issues. Section 5 details the NLP methods and techniques implemented in the *Logoscope*. Finally, Section 6 presents some sample findings obtained by using this tool.

2 General Description of the *Logoscope* Framework

The main building block of the *Logoscope* framework consists of a continuously updated lexical resource containing unknown words, their occurrences in French online press articles and further morpho-syntactic and contextual documentation. In addition our system features a query interface which allows an open access to this knowledge base.[3]

2.1 New Words are Collected in Three Stages

The *Logoscope* is built in three stages (see Figure 1). The first stage consists in the automatic detection of the unknown words. In the second stage the detected unknown words are manually validated, that is, a human expert decides which of them are genuine word creations. Finally, in the third stage these validated unknown words are documented. This is done firstly by manually adding morpho-syntactic features and information on the originating creation process. This information is automatically complemented by contextual information, typically *e.g* the journal and the paragraph the new words appeared in. But, in contrast to other similar systems (see Section 3.1), the *Logoscope* also (automatically) gives a rough approximation of what the article containing the new word is about.

In the next sections we will first give a more detailed description of each of these three stages before briefly addressing how the *Logoscope* might be used through its web interface in Section 2.5. A more detailed report of the arising NLP challenges and the methods we used to tackle them will be presented in Section 5.

[3] A simplified online version is available at the url `http://logoscope.unistra.fr`. See also Section 2.5 where we give some more details.

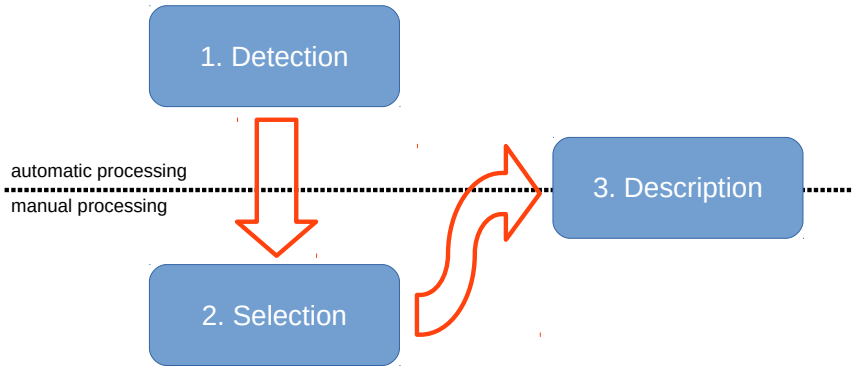


Figure 1. Processing stages in the *Logoscope*.

wid,	word,	valid,	pos,	proc,
4800,	lumbersexuel,	0,	,	,
6333,	syllogistique,	0,	,	,
7802,	ressurgissant,	0,	,	,
8068,	fantasmatiquement,	0,	,	,
4953,	franco-planétaire,	0,	,	,
6214,	consommatoire,	0,	,	,
8974,	kill-billeuse,	0,	,	,
4232,	normcore,	0,	,	,
5401,	sex-pics,	0,	,	,

(a) Sample csv (comma separated values) file listing the detected unknown words on 2015-01-02 before validation.

Occurences des mots inconnus

Occurrences de lumbersexuel

[Bing lumbersexuel] [Google lumbersexuel]

S'il ne fallait garder qu'un seul mot pour l'année 2014, c'est velu. O lumbersexuel, arrive en force et en chemise à carreaux avec des gr tout assez nerd, mais nerd des bois. Et le poil n'en finit pas de faire intimité et intriguera votre amant. Dans vingt ans, vous présenterez de bras dans toutes les couleurs possibles. E.P.

Sources : 2015-01-02#27-liberation.txt

Occurrences de syllogistique

[Bing syllogistique] [Google syllogistique]

Une rupture régalienne, un régal de rupture, celle qui a dominé l'an récit vengeur où Valérie Trierweiler défouraille son ex en mode kill l'a humiliée publiquement, elle le flingue politiquement. En vendant menti, en niant la rumeur de son aventure avec Julie Gayet, il ment

Sources : 2015-01-02#125-liberation.txt

Occurrences de ressurgissant

(b) Contexts of unknown words (as html, viewed in a browser).

Figure 2. Sample list of automatically detected unknown words and their contexts as presented to the linguist expert for validation.

2.2 Stage 1: Detection of Unknown Words

The *Logoscope* retrieves newspaper articles from several RSS feeds[4] in French on a daily basis[5]. The newspaper articles are preprocessed such that only the journalistic content is kept. The articles are then segmented into paragraphs and word forms. The resulting forms are filtered based on an exclusion list (French words found in several lexicons and corpora). They are then reordered in such a way that those words which are the most likely new word candidates appear on top, using a supervised classification method which will be described more in detail in Section 5.1.

2.3 Stage 2: Selection of the Valid Candidates

The list of unknown words produced in the previous stage is presented to a linguist expert together with the context they appeared in, for validation. Figure 2 shows the first entries

[4]Currently we collect articles from the following newspapers: *Le Monde*, *Libération*, *L'Équipe*, *Le Figaro*, *Les Echos*, *La Croix*.

[5]The average number of sources collected per day is 430 with a maximum and minimum of 968 and 28 respectively.

wid,	word,	valid,	pos,	proc,
4800,	lumbersexuel,	1,	NOM,	MORSEM,
6333,	syllogistique,	0,	,	,
7802,	ressurgissant,	0,	,	,
8068,	fantasmatiquement,	0,	,	,
4953,	franco-planétaire,	1,	ADJ,	MORSEM,
6214,	consommatoire,	1,	ADJ,	MORSEM,
8974,	kill-billeuse,	1,	NOM,	MORSEM,
4232,	normcore,	1,	NOM,	EMP,
5401,	sex-pics,	0,	,	,
		...		

Figure 3. Sample csv (comma separated values) file showing the result of the validation process for the unknown words detected on 2015-01-02.

of a list generated this way (Fig. 2a) and the corresponding contexts (Fig. 2b). The list is completed by a linguist expert who adds the following information:

a. whether the form is a genuine new word or not (column 3), *e.g.* "syllogistique" is not a new word.

b. the grammatical category of the new word, *e.g.* ADJ (adjective), NOM (noun) (column 4),

c. the type of creation process which engendered the new word (column 5), *e.g.* MORSEM (morphosemantic), EMP (borrowing).[6]

The first decision (a) is the most difficult and time consuming. For the simplest cases it can be made based on the context (the surrounding paragraph) which is presented together with the unknown form (as shown in Figure 2b), but often it also involves looking up dictionaries or the original article or performing additional investigations on the Web. It is at this point and based on this process that the linguist expert determines which new words are to enter the system and be further documented.

A typical result of this process is shown in Figure 3.

2.4 Stage 3: Linguistic Description of the Selected New Words

In the next stage, the validated new words are documented, that is they are added to our knowledge base together with the following pieces of information:

a. the grammatical category and the type of the creation process (added by the linguist expert in the previous stage, as described in Section 2.3),

b. the paragraph containing the new word,

c. further contextual information about the containing article:
   i. the journal,
   ii. the publication date,
   iii. the author (if available),
   iv. the position in the text (beginning, middle or end of the article),
   v. the thematic context, represented by the three most prominent general journalistic topics addressed in the article.

While the information in (a) has been added manually during the validation stage, the surrounding paragraph (b) and the contextual information regarding the containing article

[6] To add this information the linguist expert uses a classification introduced by Sablayrolles (2011). In Section 4 we will explain the linguistic background and motivation for this decision.

Figure 4. Documentation of the new word *lumbersexuel*.

(c) are added automatically. Whereas adding most of these items (b, ci-civ) is straightforward, automatically detecting the topics of an article requires a thematic analysis which, to our knowledge, is not proposed by any other similar tool. We will describe the methods used for its implementation in Section 5.2.

After the selection of a new word and its linguistic description, this new form is further monitored, that is we check the collected sources for subsequent occurrences and describe them in the same way as for the first occurrence. The result of this process is shown in Figure 4 for the new word *lumbersexuel*.[7] The user can see that this word first appeared in January 2015 in the journal *Libération*, that it is a common noun and that its production involved a morphosemantic process. It was then automatically detected again in April and October 2015, again in *Libération* and each time it occurred in a paragraph situated in the middle of the containing article. For each occurrence the system gives the surrounding paragraph.

As regards point (cv), the most prominent themes in the containing article are presented in different colours, as well as the characteristic words of these themes occurring in the context. In this case the result of the thematic analysis was that the articles containing *lumbersexuel* are mostly related to culture and leisure (*i.e.* they contain many terms related to culture and leisure). In Section 4 we will explain in detail why we attach particular importance to these themes. Finally the exclusion list is updated with all manually checked candidates.

2.5 The Web Interface: Using the *Logoscope*

Since the platform is targeted both at non specialists and more expert users from various scientific domains (*e.g.* linguistics and lexicography but also social sciences, politics or economics) the Web interface is designed to allow queries of a varying degree of complexity.

The most straightforward access to the knowledge base is provided by a Web interface (`http://logoscope.unistra.fr/`) where the new words can be browsed by different criteria, as for example by grammatical category, date, journal, etc. (see Figure 4).

In addition we also provide a browser-based application which allows more complex queries. Figure 5 shows a screenshot of this query interface. Here a user can use regular expressions to search for new words or query the knowledge base by several criteria (for example find new words between quotation marks, in a given period of time, occurring in articles with a given predominant topic or a combination thereof).

[7] Borrowed from the English *lumbersexual*, "A male hipster who affects a rugged, outdoorsy look, typified by plaid shirts and a full beard. " (Wiktionary, retrieved on 08-07-2016))

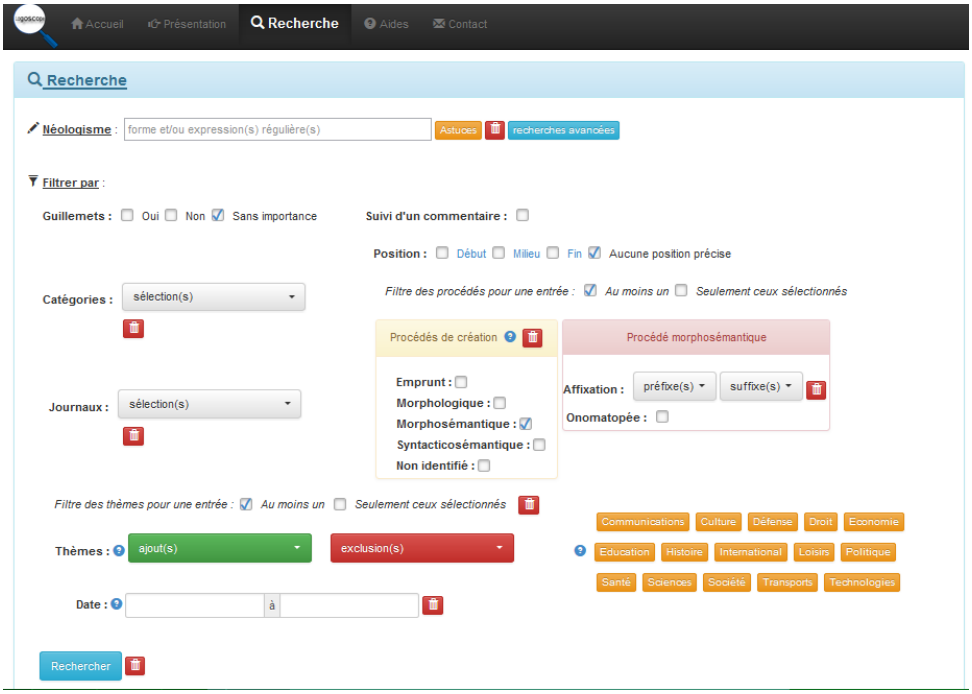


Figure 5. Query page of the *Logoscope* browser-based application.

3 Previous Work

In this section we will first present other systems devoted to the detection and documentation of new words (Section 3.1) before addressing previous work related to the two components of the *Logoscope* which mainly rely on automatic methods. The first of these two components, the automatic detection of new words, is presented in Section 3.2. In Section 3.3 we describe work related to the documentation of the detected words, focusing on an outstanding feature of the *Logoscope*, namely an approach to the automatic thematic analysis of their context.

3.1 Existing Systems and Resources for Detecting and Documenting New Words

The approaches used by systems concerned with the semi-automatic detection and observation of word creation in different languages can be analysed according to different criteria. Some of these are:

1. How the new words are detected: manually, automatically, etc.
2. The sources where new words are extracted from. These can be for example various online resources, but also documents provided by the user.
3. The information associated with the new word and presented to the end user. For most systems the goal is to elaborate a lexicographic resource. Therefore they acquire lexical information in the first place but also more or less elaborate data about the context in which the word creation occurred.

The simplest systems are databases, constituted manually and intuitively over some period of time from a collection of available documents. A typical representative of this approach is BORNEO, a database of French neologisms.[8] Another group of systems, *e.g.* POMPAMO (Ollinger and Valette, 2010), are pure extraction utilities which are fed with a text in which unknown words are detected and analysed.

[8] http://web.atilf.fr/BORNEO-Base-d-Observation-et-de

In the third more elaborate category of frameworks, the systems consist of (1) a new word detection stage based on a methodologically well defined dynamic corpus followed by (2) a documentation phase where the detected and documented new words are added to the actual linguistic knowledge base. Prominent examples of this type of approach are OBNEO (Cabré et al, 2003; Estopà, R. and Cabré Castellví, M.-T., 2004), a framework for Spanish and Catalan and NEOLOGIA (for French, Sablayrolles 2010, 2011; Cartier 2011).

These systems also differ with respect to how the new words are detected. Whereas for WORTWARTE[9] (Lemnitzer, 2010), a German neologism detection system, the acquisition process is almost entirely automated and requires little human intervention[10], for NEOLOGIA the whole documentation process is manual.

Considering lexical documentation, the amount and type of data which are retrieved and documented vary largely. Thus, for example the WORTWARTE, only records the gender, the plural mark and an ad-hoc lexical domain (*e.g.* education, health, telecommunication, environment, etc.). In contrast, OBNEO and even more so NEOLOGIA provide a very detailed lexical documentation covering (among others) the following aspects: morpho-syntax (which part of speech, grammatical sub-category *e.g. qualitative adjective*, number, gender), semantics (predicate or argument, hyper-class *e.g. human, animal, etc.*), semantic relations (synonymy, antonomy, etc.). In addition NEOLOGIA also provides a definition or gloss of the described entry and thus entirely assumes the role of lexicographer.

As already mentioned, the *Logoscope* belongs to the latter type of frameworks, based on both a new word detection and a documentation phase. In the following we will first introduce previous work concerning the detection of new words (Section 3.2). In Section 3.3 we give a theoretical analysis of the information used for the documentation of new words by the various systems, describe the items we chose for the *Logoscope* and motivate these design choices.

3.2 Methods for the Detection of New Words

Current methods for the automatic or semi-automatic identification of neologisms mainly target the coinage of new words. Some recent works have addressed semantic neologisms, with a focus on changes in part-of-speech (Janssen, 2012) or restricted case studies (Boussidan and Ploux, 2011; Reutenauer, 2012). We do not address these phenomena for the time being.

For the detection of new words, 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[11] (Lemnitzer, 2012) is an example of this exclusion list based approach.

[9] `http://www.wortwarte.de`

[10] One reason for this is, of course, that very few parameters are documented.

[11] `http://www.wortwarte.de`

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

The *Logoscope* attempts 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 appearance 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.

3.3 Features Used for the Description of New Words

As already mentioned in Section 3.1, most complex neology observation systems, namely OBNEO, NEOLOGIA and WORTWARTE, document mainly lexical information. However, in most cases they also add some information related to the context in which the new word occurred. This is not surprising since the context of occurrences is particularly important for the documentation of the new words. It not only proves and exemplifies the existence of the new word but also illustrates its meaning, especially for systems which do not provide other explanations or glosses (as is the case for the *Logoscope*). In addition to lexical specifications and the context, neology systems also provide various other information, as for example:[12]

Appearance date

Source: spoken/written, media name, competence domain[13]

Authorship: name of the author, name of the publisher, etc.[14]

Typography: dashes, quotation marks, italics, etc.

In addition, each of the more complex neology analysis systems stands out by the documentation of a particular feature: Thus NEOLOGIA is the only system specifying the position in the text, OBNEO draws the users' attention to the discourse genres and WORTWARTE proposes a basic thematic classification of the new words by manually assigning a lexical domain. In the *Logoscope* we chose to take a different approach to the thematic documentation of new words, namely by performing a thematic analysis of the surrounding text. This represents a local view on the thematic description of the new word. It allows to assess to some extent the thematic context in which the new word appeared and can be more easily implemented using automatic methods.

3.3.1 Thematic Analysis

In Natural Language Processing themes or topics[15] are generally represented in a simplified and pragmatic way, namely as a list of characteristic terms. While a single term is in most cases ambiguous, a list of such terms allows a reciprocal disambiguation and in consequence the definition of a topic.

Thematic analysis may have several goals, which we discuss in the following:

1. Detecting topics in a corpus of documents in order to determine what it is about.

2. Thematic annotation of documents based on a set of predefined topics.

[12] For a more in depth discussion please see Gérard et al (2014).

[13] OBNEO records the competence domain of the source in its corpus of spontaneous writings (*e.g.* written documents other than newspapers). Surprisingly, by "type of text" (tipus de text in Catalan) OBNEO does not mean a particular kind of text (*e.g.* law text, scientific review) but a set of publications related either to a particular professional domain (*e.g.* economy, sports, culture, computer science) or to publications of general interest (*i.e.* without a particular domain). NEOLOGIA's "domain" documents the sources using similar labels: economy, sport, society, politics, culture, etc.

[14] OBNEO also monitors a corpus of radio and television programs (Catalunya Ràdio, COM Ràdio, Televisió de Catalunya, Ràdio 9). To comply with the specificities of this corpus of spoken language, OBNEO not only documents parameters intrinsic to spoken corpora like the phonetic transcription but also the role of the speaker (*e.g.* presenter, reporter, etc.), her age and gender, dialect and mother tongue.

[15] In the following we use the terms *theme* and *topic* interchangeably.

(1) Detecting topics in a collection of documents. A technique often used for the automatic thematic analysis of a collection of documents is Topic Modeling. Topic models are probabilistic graphical models allowing to infer the topics of a collection of documents without the use of any prior knowledge or external resource. Several approaches of this type have been proposed, as for example Latent Dirichlet Allocation (LDA, Blei and Lafferty 2009). These models take a corpus of texts as input and identify a set of "latent" topics in the form of lists of words (from the corpus) characterising each theme. The model assigns to each document a finite number of topics and each topic has a probability weight. Likewise the corpus is assigned a fixed number of topics with different probability weights. However the inferred topics are not labelled and therefore an important analysis effort is necessary to name them. It is this kind of approach we used for the thematic analysis in the *Logoscope*.

(2) Thematic annotation of documents. When a thematic resource is available, documents can be annotated with respect to this resource. In this case the granularity of the thematic analysis depends on the granularity of the resource used for the annotation. Beust (2002) proposes a thematic markup tool called ThemeEditor.[16] This utility characterises each theme or topic using manually defined lists of terms. In a text given as input, the terms belonging to each theme are then highlighted using different colours. The words which characterize a topic are similar to the keywords or key expressions corresponding to the most important subjects of a document. While in many studies keywords or expressions are extracted based on purely statistic methods, Medelyan and Witten (2008) introduce a method for automatic indexation where the keywords assigned to the documents are controlled by aligning them to an existing thesaurus. Medelyan and Witten first identify which terms from the thesaurus occur in the text. Then they use a supervised model to select the most characteristic terms. The model is trained using various features: the `tf.idf`[17] score, the position of the first occurence of a term in the text, the number of words making up the term, semantic relation to other candidate terms in the document. Being supervised, a set of mannually indexed documents are also needed. This number has been estimated to 50–100, which reportedly only required a limited effort.

***Logoscope*'s thematic analysis** For the thematic analysis implemented in the *Logoscope* we first automatically retrieved a set of general journalistic topics using topic modeling (as in 1). These were then manually labeled and remodeled into a resource (collection of themes, which in turn are represented by lists of terms) which could then be used to thematically mark up the articles containing the new words (based on methods similar to 2).

4 Linguistic Principles of the Logoscope

In this section we discuss the linguistic motivations which guided our choices concerning the documentation of the new French words. We specify which data we decided to add to the *Logoscope* knowledge base for each new word and explain the reasons for these decisions.

In general, a neology observation system[18] should not only document a large number of lexical features (*e.g.* grammatical category, verb valency, production type, etc.) but also contextual parameters (*e.g.* appearance date, context, quotation marks, etc.) to characterise the words (cf. Gérard et al 2014).

Lexical features: a minimalistic description. In the *Logoscope* we deliberately restricted the number of lexical features to the two which we considered essential, namely the grammatical category and the production process. This minimalistic approach is justified by the following considerations. First, the *Logoscope* is not meant to offer a lexicographic

[16] `https://beust.users.greyc.fr/ThemeEd`
[17] Term Frequency Inverse Document Frequency
[18] Cf. the systems described in Section 3.1, as for example WORTWARTE, OBNEO and NEOLOGIA.

(Meta-)category	Type of lexical creation process
morpho-semantic	affixation inflection composition(s) onomatopea paronymy or approximate homonymy
syntactico-semantic	conversion(s) lexical and syntactic combinations metaphor metonymy
morphologic	truncation abbreviation or using only initials
loan	identical borrowing assimilation of loan words

Table 1. Lexical categories documented by *The Logoscope*.

interpretation: we deliberately do not propose any word definition. Second, this lexical information is nonetheless sufficient as a basis for further investigation by the user according to their particular centres of interest (journalistic, linguistic, sociological, etc., see Section 6). Last but not least, in this way the validation time is minimised, which is crucial considering the relatively large number of daily incurring forms needing validation.[19]

Regarding the grammatical category, we were concerned with choosing a small number of well known part of speech tags but which should ideally cover all potential newly created words. This resulted in a set of currently ten grammatical categories: adjective, adverb, gerund, interjection, common noun, proper noun, past participle, present participle, pronoun, verb.

To best document the second lexical feature, the type of process responsible for the word creation, we follow a categorisation proposed in Sablayrolles (2011) which is outlined in Table 1. Ultimately we use the meta category, the first column in Table 1 to describe the new words ignoring the more specific description of the creation procedure shown in the second column. Thus for example, *songwriteuse*[20] will simply be described as a *common noun* and *loan+morpho-semantic*, but not as an *assimilation of loan word+affixation*. This way the annotator may gain time and minimise the error rate but nonetheless provide relevant information allowing a subsequent more thorough lexicographic research.

Contextual parameters: a maximalistic description. One of the specificities of the *Logoscope* is the automatic documentation of contextual parameters. This approach stems from the fact that each lexical creation is dependent on the communicational situation it occurs in and is tightly interlinked with the text in which it is read and understood. Moreover, not only the sense of a new word, but also the associated linguistic and extralinguistic characteristics, its rhetoric function and even the generated mental representation strongly depend on the particular circumstance in which the interpretative activity happens (cf. Gérard 2014; Dal and Namer 2016).

Our approach belongs in the larger field of text linguistics which places texts (rather than sentences) at the centre of the study of communication systems (cf. Gérard 2011; Solé 2002). This research direction has been strongly advanced from the 1980s on in particular by researchers in Germanic linguistics (reflected for example in Peschel 2002). At present it is notably pursued in studies about particular textual genres, such as novels, blogs, etc. (*e.g.* Siebold 2000; Gérard and Lacoste 2016).

From a more practical point of view, for our description purpose we consider that there are two kinds of contextual parameters, namely textual and discursive variables. As shown in Figure 6, the former are intrinsically tied to the semiotic material of the text while the

[19] Currently the system collects ≈50 forms on average, after an initial "burn-in" period of several weeks where it daily collected more than 100 forms.

[20] « L'histoire officielle raconte que Courtney Barnett a 26 ans et qu'elle fut serveuse à Melbourne. Pêchue, acidulée, tonique, c'est surtout la **songwriteuse** de ce printemps » (Libération, 23/03/2015).

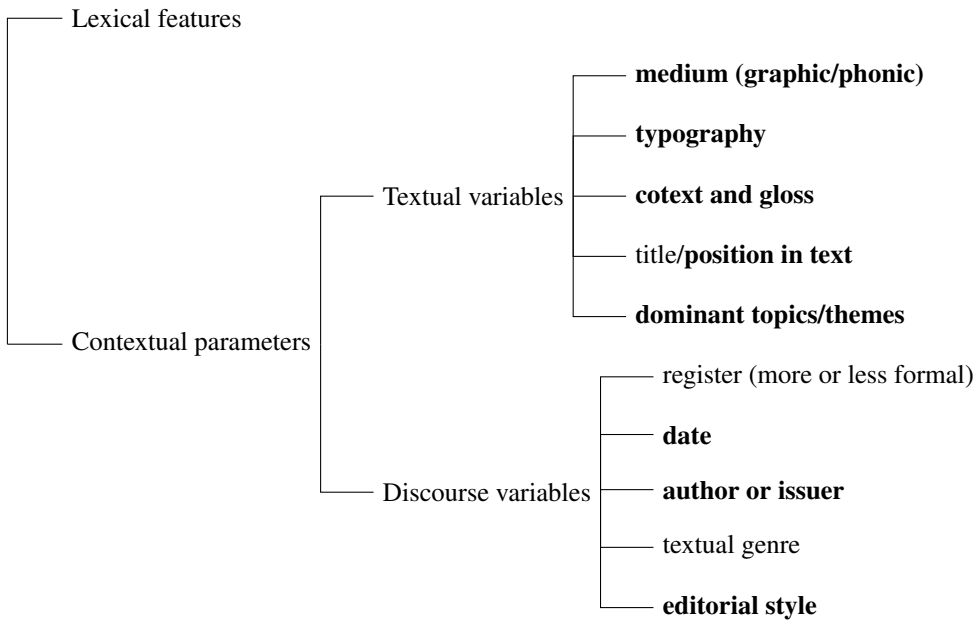


Figure 6. Contextual parameters for the description of new words. The parameters set in boldface are those documented in the *Logoscope* framework.

latter rather reflect the communication situation and the currently valid socio-discursive norms. In addition to this typology the figure also indicates in boldface those features which are effectively collected and stored in the *Logoscope* knowledge base.

Only the title, the register and the journalistic genre (*e.g.* editorial, opinion, portrait, interview, etc.) cannot be documented automatically. The reasons for this are very diverse. While they are of purely technical nature for the *title* – the title is marked up in many different ways, even for the same journal or site – they are more fundamental for the automatic detection of the *journalistic genre*[21] and *register* or *code*[22]. To our knowledge currently there are no computational methods sufficiently sophisticated to allow for the automatic detection of these variables.

The thematic analysis. We finish this section by a brief introduction of *Thematic*, a tool which is dedicated to the thematic analysis of newspaper articles and thus allows the automatic detection of their dominant themes. The results of this thematic analysis are used to fill in the thematic component of the contextual variables. They can also be visualised through a thematic colouring as shown in Figure 7.

Figure 4 illustrates how these results are used in the documentation of the new words. It shows that in most cases the articles containing the new word *lumbersexuel* were related to the topics of culture and leisure (green/brown colour).

We will give more details about the theoretical and technical background in Section 5.2.

5 Implementation: NLP Techniques Employed

This section describes prominent NLP methods we used in the *Logoscope* and shows how they were implemented to achieve the largely automatic detection and documentation of the

[21] To our knowledge *OBNEO* is the only system documenting, manually, the feature of discourse genre by specifying the genre for the radio or television broadcasts in their corpus of spoken language.

[22] Some systems annotate new words as "spoken" or "written". Thus in *BORNEO* words as *craignoss*, *trip musique*, *à chier* or *méga-sale* are assigned an "oral" quality. However, we consider that the spoken respectively written quality is not only engendered by the different media but must be viewed in the larger communicational context: The speakers realise that their communicational behaviour must be appropriate to the present setting and they adapt to the situation by adopting a "style" which is rather free, less formal or affective or on the contrary more formal, academic or reserved (Wulf Oesterreicher and Peter Koch 2001).

Résultat de l'analyse de "LeMonde_Pape.txt" :

-- Le pape condamne la " persécution brutale " des " minorités " La préoccupation du pape pour le sort des chrétiens d ' Orient s ' est exprimée tout au long de cette semaine de Noël . Lors de la bénédiction urbi et orbi qu ' il a prononcée à midi , jeudi 25 décembre , place Saint-Pierre à Rome , à l ' occasion de la fête de la nativité , François a dénoncé les " persécutions brutales " dont sont victimes les chrétiens d ' Irak et de Syrie , avec ceux " qui appartiennent à d ' autres groupes ethniques et religieux " . Il a évoqué " les nombreuses personnes déplacées , dispersées et réfugiées […] de la région et du monde entier " et a demandé qu ' elles " puissent recevoir les aides humanitaires nécessaires pour survivre à la rigueur de l ' hiver et revenir dans leur pays " . Lundi , le pape argentin avait adressé une lettre aux chrétiens du Moyen-Orient pour les " encourager " et leur dire " combien [leur] présence et [leur] mission sont précieuses en cette terre " où " est né et où s ' est répandu le christianisme " . Comme à son habitude , il n ' y citait pas nommément l ' État islamique mais il y dénonçait une " organisation terroriste " qui " commet toutes sortes d ' abus et de pratiques indignes de l ' homme , en frappant de manière particulière certains d ' entre vous qui ont été chassés de façon brutale de leurs propres terres , où les chrétiens sont présents depuis les temps apostoliques " . TROP D ' ENFANTS " VICTIMES D ' ABUS ET EXPLOITÉS " Jeudi 25 décembre , le pape François a prononcé le deuxième discours de Noël de son pontificat . Jeudi , devant les dizaines de milliers de personnes assemblées devant la basilique Saint-Pierre , le pape a demandé un large " soutien aux efforts de ceux qui s ' engagent efficacement pour le dialogue entre Israéliens et Palestiniens " . Il a aussi mentionné l ' Ukraine , le Nigeria , où " à nouveau du sang est versé et trop de personnes sont injustement soustraites à l ' affection de leurs proches et tenues en otage ou massacrées " , ainsi que d ' autres pays africains : la Libye , le Soudan du Sud , la République centrafricaine et la République démocratique du Congo . Dans une bénédiction où , comme il l ' a souligné , étaient présentes " tant de larmes en ce jour de Noël " , le souverain pontife a également dénoncé les mauvais traitements infligés à des enfants , victimes de violences , de trafics , de traite des personnes ou contraints à être soldats . Il y a trop d ' enfants " victimes d ' abus et exploités sous nos propres yeux et avec notre silence complice " , a -t-il déclaré . Il a cité " les enfants massacrés sous les bombardements , y compris là où est né le fils de Dieu " , c ' est-à-dire en terre sainte , et leur " silence impuissant qui crie sous l ' épée " . Il a aussi mentionné les enfants tués à Peshawar , au Pakistan , dans l ' attentat contre une école la semaine dernière , ainsi que ceux victimes du virus Ebola au Liberia , en Sierra Leone et en Guinée . Dans une référence à l ' avortement , il a aussi cité les " enfants tués avant de voir la lumière " .

Thèmes Spécifiques :

Thèmes Principaux	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Mythologie, Religion et Croyances	24	14	22.58	94.10	1.37%	1.96%
Conflits Armés Contemporains	18	13	9.28	51.57	4.63%	1.47%
Afrique	8	8	8.00	100.00	1.97%	0.65%
Thèmes Secondaires	**Tokens**	**Types**	**Poids du Thème**	**Poids/Tokens du Thème**	**% mot colorié sur le total de mot du Thème**	**% mot colorié sur le total de mot de l'article**
Moyen-Orient	7	7	7.00	100.00	3.32%	0.57%
Démographie	10	5	6.00	60.00	2.15%	0.81%
Agriculture	6	5	2.83	47.22	2.28%	0.49%
Médecine	5	5	3.25	65.00	0.82%	0.41%
Justice et Législation	6	4	3.00	50.00	0.42%	0.49%
Souffrances Psychologiques et Physiques	4	4	3.00	75.00	0.84%	0.33%
Militaire	4	4	2.33	58.33	0.27%	0.33%
Discrimination et Égalité Sociale	4	3	4.00	100.00	1.11%	0.33%
Musique	4	3	2.58	64.58	0.21%	0.33%
État et Gouvernement	4	3	2.00	50.00	0.62%	0.33%
Immigration	3	2	3.00	100.00	1.80%	0.24%
Amériques	3	2	3.00	100.00	1.05%	0.24%
Cinéma et Théâtre	2	2	1.33	66.67	0.22%	0.16%
Antiquité (-3000 > 496)	2	2	1.00	50.00	0.17%	0.16%

Thèmes Génériques :

Thèmes Génériques	Tokens	Types	Poids du Thème	Poids/Tokens du Thème	% mot colorié sur le total de mot du Thème	% mot colorié sur le total de mot de l'article
Société	38	22	32.58	85.75	0.85%	3.10%
Défense	22	17	11.62	52.80	0.81%	1.79%
International	18	17	18.00	100.00	0.99%	1.47%
Santé	9	9	6.25	69.44	0.74%	0.73%
Politique	7	5	5.00	71.43	0.71%	0.57%
Culture	6	5	3.92	65.28	0.07%	0.49%
Économie	6	5	2.83	47.22	0.45%	0.49%
Droit	6	4	3.00	50.00	0.37%	0.49%
Histoire	2	2	1.00	50.00	0.10%	0.16%

Figure 7. Example of a thematic analysis on a *Le Monde* article (25-12-2015). According to this analysis, the text is principally about a societal topic, but it also employs many words related to the theme of defense in an international context. Within the main topic, the most important subject was found to be about "mythologies, religion and believes". In our view, this analysis gives a reasonably good idea about the overall subject of the text, even if it is far from showing anything about the exact content.

new words. The first set of such NLP techniques, presented in Section 5.1, is used in the detection of the new words. Section 5.2 is about our implementation of a thematic analysis which permits a novel approach to the documentation of the textual context of new words.

5.1 Detection of Unknown Words

As discussed in Section 2.2, most methods addressing the detection of formal neologisms are based on exclusion lists. The *Logoscope* also relies on exclusion lists but in addition the (necessary) manual validation is alleviated using a statistical machine learning method which ranks the words not in the exclusion list with respect to the probability that they are interesting neology candidates. Here we present a brief overview of our approach, as it already has been described in detail in Falk et al (2014).

As mentioned in Section 2.2, the *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 new words. For example, Table 2 shows the most frequent unknown words[23] 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. The method we developed ranks these detected unknown forms with respect to

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

the probability that they are interesting neology candidates. Thus, improbable candidates as *e.g. candiudature* or *lmd* would be removed to the bottom of the list whereas more probable candidates as *geniculatus*[24] are put at the top of the list.

Our approach is a classical machine learning (supervised classification) approach: We use the extracted forms (and various characteristics or features thereof) and the annotations (valid new word or not) by a linguist expert. Based on this data the system learns a model such that each extracted form is assigned the probability of its being a valid new word. The word forms are then ranked according to this probability. Thus, hopefully, in the validation step, the linguist expert will be presented with the most probable valid new words first.

In our approach we experimented with three types of features: form related, morpho-lexical and thematic or contextual features. Our experiments highlighted the importance of the thematic features, which, to our knowledge have not yet been used in this kind of application. In addition, these features represent a way to access and document the semantic context of the new words.

Features. We explored the effect on the classification of three groups of features: *formal* features, *morpho-lexical* features and *thematic* features, an overview of which is presented in Table 3. The formal 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 such 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[25]. We

[23] collected on July 12, 2013

[24] *geniculatus* was not found to be a new word, but it is still a more probable candidate as most other forms in this list

[25] We used the `Lingua::Identify` perl script to this end: `http://search.cpan.org/~ambs/Lingua-Identify-0.56/lib/Lingua/Identify.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
Levenshtein distance to suggestions from spell-checker (*aspell*) or very large value Does the form contain other known forms? (Aho and Corasick 1975 algorithm)

(b) Morpho-Lexical features.

Topics	
10 topics extracted from the Le Monde Corpus	
Topic features: 10 features, proportion of each topic in each document	
Documents	Feature value
articles containing form	max. topic proportion
concatenation of paragraphs containing form	topic proportion
Newspaper: The newspaper(s) the form appeared in.	

(c) Contextual or thematic features.

Table 3. Overview of features used in the classification task.

also assess the probability that the form might be a spelling error by using the aspell tool[26]. Finally, 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[27]. 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 explore the influence of thematic features on the automatic identification of the new words (Table 3c). Our hypothesis is that these features supply interesting additional information not provided by form related and morpho-lexical features. A first obvious feature describing the context is the *Newspaper* feature (the newspaper where the unknown form was found). In addition, we attempt to capture the thematic context of the text containing the unknown form using a technique called *topic modeling* Steyvers and Griffiths (2007). Topic models (cf. Section 3.3.1) 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, 2013)) show that in general tens or hundreds of thousands of documents are needed

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

[27] This group of features is derived from the results of the Aho and Corasick (1975) string matching algorithm which suggests a list of known forms present in the unknown form at hand

formal, morpho-lex, thematic					*formal, morpho-lex*			
class	**Prec**	**Rec**	**F**	**corr.**	**Prec**	**Rec**	**F**	**corr.**
pos	0.181	0.827	0.297		**0.192**	0.778	0.308	
both	0.868	0.548	0.625	67	0.864	0.597	0.669	63
formal, thematic					*formal*			
class	**Prec**	**Rec**	**F**	**corr.**	**Prec**	**Rec**	**F**	**corr.**
pos	0.160	0.531	**0.346**		0.190	0.481	0.273	
both	0.826	0.625	0.693	43	0.832	0.704	**0.752**	39
morpho-lex					*thematic*			
class	**Prec**	**Rec**	**F**	**corr.**	**Prec**	**Rec**	**F**	**corr.**
pos	0.132	0.827	0.227		0.129	**0.889**	0.225	
both	0.836	0.350	0.415	67	0.844	0.295	0.338	**72**

morpho-lex, thematic				
class	**Prec**	**Rec**	**F**	**corr.**
pos	0.136	0.877	0.236	
both	0.851	0.345	0.404	71

Table 4. Classification results. In green (blue) best respectively worst F-measure and true positive results. Best balanced results highlighted in orange. Overall best precision, recall and F-measure results are set in bold face.

for a thorough thematic analysis of this kind and that the number of extracted topics is between 100 and 300. In our preliminary experiment we collected 4,755 articles from the newspapers shown in Table 3c and restricted the number of extracted topics to 10.

The features for each unknown word are obtained by applying topic modeling to two types of context. The first is obtained by concatenating all the sentences containing the unknown form. We expect the result of the topic analysis on this concatenated context to represent the weight of each topic in the closer phrasal context of the unknown word. The second type of context we use are the articles containing the unknown forms. We apply the topic analysis to each article containing the unknown word and associate the unknown word with an average of each of the topic proportions over the articles. We expect these features to represent the predominant topics of the articles containing the unknown forms.

Classification method. Our new word detection problem is casted into a supervised classification problem in the most straightforward way: the validated new words are considered as positive examples and the remaining unknown forms as the negative examples in the training data. We accounted for the strongly imbalanced data[28] by oversampling the positive class[29]. We then used the SVM classifier as implemented in LibSVM Chang and Lin (2011)[30] to perform the classification. More specifically, we produced 7 classifications (see Table 4), one for each combination of the three groups of features described earlier: the *formal*, *morpho-lexical* and *thematic* feature groups. We then evaluated these classifications in 10-fold cross-validation by looking at precision, recall and F-measure for the positive class and on the average over both the positive and negative class. In addition we also report the number of validated new words (the true positives).

The results are shown in Table 4. The best F-measure (highlighted in green) for the global classification task was obtained using the *formal* features, but in this setting the smallest number of validated new words could be identified (in blue). The highest number of validated new words could be identified using the *thematic* set of features but in this case the global F-measure was comparably low. The best balance between global F-measure and detected new words was obtained using the *formal, morpho-lex, thematic* feature combina-

[28] 81 positive examples vs. 611 negative examples

[29] 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 WekaHall et al (2009) cost sensitive classifiers to achieve this.

[30] We used an exponential kernel with cost 1 and $\gamma = 0$

ultra-présent (−)	crypto-fascisme (−)	anti-défilé (−)
Etat-département (−)	semi-itinérants (−)	pro-MDC (−)
anti-alcoolisme (−)	mini-Internationale (−)	anti-monégasque (−)
pagano-satanisme (+)	neo-retraité (+)	entraîneur-athlète (−)
watts-étalons (−)	écarts-types (−)	néonicotinoides (−)
auto-diagnostiqués (−)	agroécologiste (+)	...

Table 5. Unknown words ranked by SVM probability. Classification obtained with *form, lex, theme* features. In parentheses: if validated (+) or not.

tion. Overall 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 5 shows a possible reordering produced by our system. They also suggest that the features used in our experiments were sufficiently powerful to support this classification scenario outweighing the unbalanced data and the difficulty of the classification task.

Groups of features. The thematic features played an important role in the classification, since the classification model based on combinations involving the thematic features achieved good results. Thus the classification using the thematic features had the best recall score for the positive class and the F-score for the positive class was highest for the combination of thematic and formal features. First this confirms our intuition that the context is helpful at the detection of new words, a finding 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 Section 4). Second, this highlights the benefit of using topic modeling for an additional representation of thematic content. Indeed, this way it is possible to assess the semantic content of a wider textual context than the more limited co-occurence windows which are used ordinarily. This aspect, as shown in Section 3 is rarely taken into account, theoretically or practically, by recent neologism detection utilities.

Qualitative discussion. For a qualitative analysis we applied the classification models based on the *formal*, *morpho-lex*, *thematic* and *formal, morpho-lex, thematic* feature groups on our data and examined the number of correctly identified new words and, for each group, the five best scoring unknown words and the five best scoring validated new words (Table 6). This allows to better tease out the effect of the various types of features on the selection of valid new words.

First we observe that the *formal* features help identify particularly long new words, and those containing a hyphen (first line).

The second line shows that the words scoring best with the *morpho-lex* features are mainly compositions (with or without hyphen) or contain a prefix (*anti, non*).

With respect to the contextual (*thematic*) features we observe on the positive side that they permit the detection of new words with no prominent property (*agnélise, retricoté*), but on the negative side these thematic features seem to favour the selection of words which are not plausible considering traditional formation rules for French word forms (*e.g. schlopp, gesagt*). A closer look at the new words detected through the *thematic* features but not via *morpho-lex* and *formal* features confirmed their ability to select new words with less characteristic forms. Thus, some new words identified by the *thematic* classifier, but not by the *morpho-lex* and *formal* classifiers are: *accrobranches, caricatureurs, conflicté, frenemies, instinctivores* ....

In the current version of the *Logoscope*, the unknown words are ranked using an SVM classifier based on the *formal, morpho-lex, thematic* feature combination. Since this study confirmed the importance of the thematic features we produced a more enhanced topic model of 100 topics based on the Le Monde corpus and use this model to compute the thematic features. This topic model is also used at the assessment of the thematic context of the new words (and in consequence at their documentation) and will be described more in detail in the next section.

Features	#neos	top 5 valid new words	top 5 (new word?)
form	37	supermédiateur, doublevédoublevédoublevé, auto-diagnostiqués, néo-célibataires, surmonétisation	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, parte-nadversaires, 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 6. Top 5 predictions when applying the model.

5.2 Thematic Analysis

The goal of our thematic analysis is to give a general idea about the thematic context in which word creation occurs. We illustrate this goal by an example produced by the *Logoscope* and shown in Figure 7. The presented article gives an account of several statements issued by the Pope at a Christmas celebration. The thematic analysis found that the text was mainly about a societal topic and showed that it also deals with some issues related to the theme of defense in an international context. We consider that this gives a reasonably good idea about the general thematic background of the subject addressed in the text, even though it is far from describing the exact content.

Ideally for this type of analysis one would need, in the first place, a register as complete as possible, of journalistic topics or themes (*e.g. Société/Society* or *Défense/Defence* in the example). Unfortunately, to our knowledge such a register is not available. Typically newspaper outlets do assign their articles to particular categories, but the resulting categorisation is neither systematic, nor uniform and can not be expected to be complete. We therefore first developed a method to obtain such a register of general journalistic topics, reflecting the themes a newspaper article might be about.

Based on this set of journalistic topics, we then developed a method to automatically associate a newspaper article to those topics which best reflect its thematic content. The result is a thematic analysis as shown in Figure 7. Later, in *Logoscope*'s documentation stage, the topics/themes which were found to be prevalent for that article are associated to the article, and by proxy to the new words it contains.

In the following we first describe the acquisition of the general journalistic topics and then our method to assess the thematic content of a newspaper article.

Acquiring general journalistic topics. As already discussed in Section 3, one of the few eligible methods to automatically represent and assess thematic content is topic modeling (Blei and Lafferty (2009)). In a nutshell, in topic modeling, documents are viewed as a mixture of topics, where a topic is a probability distribution over words. Standard statistical techniques allow to invert this process and infer the topics (in terms of lists of words) that were responsible for generating the given collection of documents. The learned topic model can then be applied to an unseen document and the thematic content of this document can thus be estimated in terms of the inferred topics.

groupe	0.05867	droit	0.02779	jean	0.23761
société	0.02616	conseil	0.02254	pierre	0.09342
capital	0.01642	loi	0.01790	claude	0.05313
actionnaire	0.01381	pouvoir	0.01520	jacques	0.05284
filiale	0.01078	commission	0.01475	marier	0.05010
affaires	0.00926	rapport	0.00992	louis	0.03668
entreprise	0.00912	cas	0.00952	alain	0.03400
franc	0.00907	devoir	0.00805	marc	0.02243
milliard	0.00860	texte	0.00794	saint	0.01784

(a) Topic 0 **(b)** Topic 2 **(c)** Topic 19

Figure 8. Words with highest probabilities in some topics acquired by applying Latent Dirichlet Allocation to the Le Monde corpus. Topic 0 and Topic 2 are examples of meaningful topics in our setting, which could be labeled easily. In contrast Topic 19 shows a topic which could hardly be used for representing the content of newspaper articles.

We acquired the topics from the Le Monde corpus, a collection of more than 900 000 newspaper articles from the French journal *Le Monde*, dating from 1987 to 2006[31]. Because of the nature of this corpus, we expect the resulting topic model to reflect the content of the newspaper articles collected by the *Logoscope* fairly well. To this corpus we applied Latent Dirichlet Allocation (LDA), a probabilistic graphical model and generated a set of 100 general journalistic topics. These topics are represented as probability distributions over words. Figure 8 shows the words with highest probability for some of these topics. We found that many of these topics were meaningful, in the sense that it seemed doable to figure out the "latent" theme they represent and give it a suggestive label. For example, one could say that Topic 0 in Figure 8a is about economics and Topic 2 in Figure 8b about law or justice. For others it was obvious that the underlying theme was not interesting. An example for this is Topic 19 in Figure 8c which apparently is mainly a collection of male first names.[32]

Hence we could not use the topics directly but had to remodel them. For this we selected the most interesting topics (currently 71). In these topics we only kept the significant terms and also added some terms if necessary. This had to be done manually and required a considerable effort. It resulted in a collection of 71 general journalistic themes which are meaningful to human readers and reflect the content of the journals we are dealing with.

Topic models generated by LDA only can easily be used in a subsequent step to infer, for an unseen text, a probability distribution of the learned topics. Because in the *Logoscope* we modified the words in the topics, this computationally cheap automatic inference is no longer available.

Figure 9 shows two examples of resulting themes.[33] In contrast to the topics, we also no longer dispose of the probability weights and therefore all terms are equally important for the thematic analysis of a text.

Assessing thematic content. Based on the themes developed this way we implemented our thematic analysis as follows. Input to our method are the 71 general journalistic themes and a text, typically a newspaper article. The result is on one hand a thematic colouring of the text as shown in Figure 7. On the other hand the text is also associated with the

[31] We used the gensim toolkit to compute the topic model (Řehůřek and Sojka (2010), `https://radimrehurek.com/gensim/`).

[32] To our knowledge currently there is no principled way to automatically "label" the topic lists produced by this method.

[33] It also shows that the terms in the themes comprise the POS (grammatical category) and that we also use compound terms, in an effort to increase accuracy and avoid lexical ambiguity. We obtained the grammatical categories by preprocessing the texts with TreeTagger and adding the corresponding grammatical categories to the terms in the themes. The themes also contain compound terms, which are often highly characteristic of a particular topic. Thanks to lemmatisation and POS tagging it is possible and technically not to expensive to match them in the articles.

eurogroupe-nc
société-nc général-adj
capital-nc
capitalisme-nc
capitaliste-adj
capitalistique-adj
actionnaire-adj
actionnaire-nc
actionnarial-adj

(a) Economie-Finance (economy and finance).

droit-nc
avoir-v droit-nc
chambre-nc du-prp conseil-nc
conseils-nc
loi-nc
loi-nc carrez-nc
pouvoir-nc
commission-nc
commission-nc européen-adj

(b) Droit-JusticeLegislation (law, justice, legislation)

Figure 9. Examples of resulting themes: Economie-Finance (economy and finance) and Droit-JusticeLegislation (law, justice, legislation).

most relevant themes – this is what we presume the text is about[34].

We determine the most relevant themes by looking at the terms the themes and the text have in common. The more terms shared by a theme and the text, the more relevant we consider the theme for this text[35].

To wrap up, with this flavour of thematic analysis we found a viable way to assess the thematic context of new words and to associate them with labeled themes which are more meaningful for human readers than the topics resulting from topic modeling.

6 The *Logoscope* at Work: Linguistic and Extra-linguistic Applications

The aim of this section is to give a general idea about how the *Logoscope* could be used in an analysis of word creation. We first present some findings about word creation in French newspaper texts which we could obtain using the *Logoscope*. We then illustrate its use for classical linguistic and lexicographic analyses but also for some basic observations pertaining to the realm of social and/or political sciences.

At the time of writing (end of 2017), the *Logoscope* data base contained 1126 new words together with the almost 15 000 paragraphs in which they occured. Most occurences appeared in the journal "Les Echos", followed by "Le Figaro" and "Libération". However, to be more meaningful, these numbers still need to be related to the total number of paragraphs collected for each journal.

In almost 60% of instances the new words were used as common nouns and the second most frequent category were adjectives (almost 20%). Interestingly, almost 18% of instances were used as proper nouns. Some examples are various names used to write about newly built administrative units, as *alsace-champagne-ardenne-lorraine* or *haut-de-france* (see below for some more discussions). Others are derived from proper names by affixation (*dieselgate, ex-erdf*) but there are also genuine innovations as for example *Boxit*, a contraction of *Boris* and *exit*.

Regarding the creation process, most new words (almost 60%) emerged from a morpho-semantic process and almost 40% are also loan words. Some sample new words which were created by borrowing and/or a morpho-semantic process are shown in the following.

1. Pour utiliser une monnaie virtuelle, il faut disposer d'un portefeuille électronique, qui stocke bitcoins, **ethers** ou ripples. Les portefeuilles sont disponibles sur les smartphones via des applications ou sur les ordinateurs via des sites dédiés.

2. Lycéenne le jour, **startuppeuse** la nuit. A 17 ans, Philippine Dolbeau développe New School depuis déjà un an et demi. Ce système vise à remplacer et à automatiser les

[34] In the example in Figure 7 these are *Mythologie, Réligion et Croyance* (mythology, religion, religious faith), *Conflits Armés Contemporains* (contemporary armed conflicts) and *Afrique* (Africa).

[35] To compute the term overlap efficiently we use again the Aho and Corasick 1975 algorithqm.

cahiers d' appel à l'école. Lauréate de plusieurs prix, elle est notamment suivie par Apple.

3. Il existe des « professeurs », des « docteurs » et des « restaurateurs », mais pourquoi pas des « **professeuses** », « docteuses » ou « restaurateuses » ? Claude Duneton (1935-2012) analysait l'impossible féminisation des mots en « -eur ». Le Figaro vous fait redécouvrir sa chronique.

4. Va-t-on assister à la naissance d'un **assado-poutinisme**, ultime avatar d'une politique étrangère décidément incapable de renouer avec la mesure et l'intelligence de notre nouveau monde ? (. . .)

In the first example (1) the new word *ethers* is a pure loan word whereas in the second example (2), *startuppeuse* the borrowing is combined with a morpho-semantic process (affixation). Finally the last two examples (3, 3) show new words (*professeuses*, *assado-poutinisme*) which emerged through a morpho-semantic process only (affixation and composition resp.). Example 3 is also interesting from a lexicographic point of view because it points to the fact that in French there are few feminisations of words in "-eur", therefore those which are created are particularly outstanding. These examples show how the *Logoscope* could be used for some traditional lexicographic investigations, as for example observing which are the most productive word-creation processes, or which affixes are most frequently used in the word creation process.

The position where the new words appear gives some hints about some journalistic practices. In most cases (54%) they appear in the middle of the newspaper articles, which is not surprising, since this is the main body of the texts. However an important portion (30%) of the instances are used in the last three paragraphs, a relatively small part of the article, suggesting that a common praxis was to "drive home" a point by using a surprising new word. In contrast, new words were relatively rarely used in introductory paragraphs (16.5%)[36] – apparently they were not found to have a sufficiently attention-grabbing effect.

6.1 New Words and Textuality: going deeper in the study of lexical creativity

The *Logoscope* naturally allows for a traditional lexicographic analysis of new words. But due to its specificity of taking into account the textual context of word creation it also permits to study these new words there where they are used: in the surrounding texts. For example, as our brief discussion above shows, it allows empirical observations regarding journalistic style.

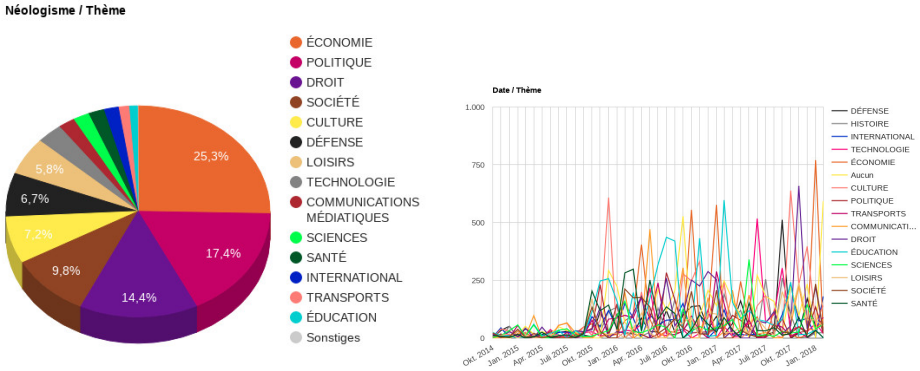
However, while other systems exist which allow to observe the interaction of word creation with journalistic style, the *Logoscope* also makes possible more systematic empirical studies at a larger scale due to the integrated more extended contextual documentation. More specifically, to our knowledge there are no other new word detection systems which collect data about the thematic context where the word creation occurred.

The following very sketchy analysis of word creation in terms of its larger thematic environment is meant to give an idea of how this specificity of the *Logoscope* could be used.

First, Figure 10a gives an overview of the topics of the articles where most new words were used. Even though to be more meaningfull this should be related to the overall distribution of articles over the topics, we think that this visualisation nevertheless allows a first rough estimation of the thematic context where most new words were used. We see that new words mostly appeared in articles dealing with economics (25%) followed by articles about politics (17%) and law (14%). Figure 10b shows in what thematic contexts most new words were used over time. Thus in January 2016 most collected new words were used in articles with a cultural background, whereas in February 2017 the articles containing the new words were mostly related to culture. "Zooming into" the time frame of January 2016 we can find some examples of new words used in a cultural thematic context:[37]

[36] The introductory paragraphs include the title and chapeau because it was not possible to distinguish them from the main body paragraphs.

[37] New words are shown in bold-face, we don't show the entire paragraph.



(a) Topics of newspaper articles where most new words were used.

(b) Occurrences of new words in articles about different topics over time.

Figure 10. Observing the thematic context in which word creation occurs. The figures are screenshots obtained from the online version of the *Logoscope* (`http://logoscope.unistra.fr/topic/`).

- Oui, beaucoup ont espéré qu'on se fasse tuer. **tu-er** », poursuit-il, en rappelant la fragilité du journal (...)[38]
- ...le FFZero1 concept apparaît comme un **hypercar** très inspiré de la Batmobile de Batman ...

Finally we give some examples of new words used in articles related (among others) to education, from February 2017:

- En matière de cibles publicitaires, les **millennials** (20-30 ans) ont fait dernièrement beaucoup parler d'eux.
- Michel Agier : « Repenser l' engagement citoyen est le seul moyen d'éviter l'**encampement** des migrants »
- David Forge, jeune agriculteur et **youtubeur** le 30 Janvier 2017 à Saint Senoch .

6.2 Spotting and Documenting New Realia in French Society (2014-2016)

In this last section, we would like to briefly show how the *Logoscope* contributes, beyond lexicologic and lexicographic issues (traditionally associated with new words), to spot and document the emergence of new French realia, that is realities which are specific to contemporary French society.

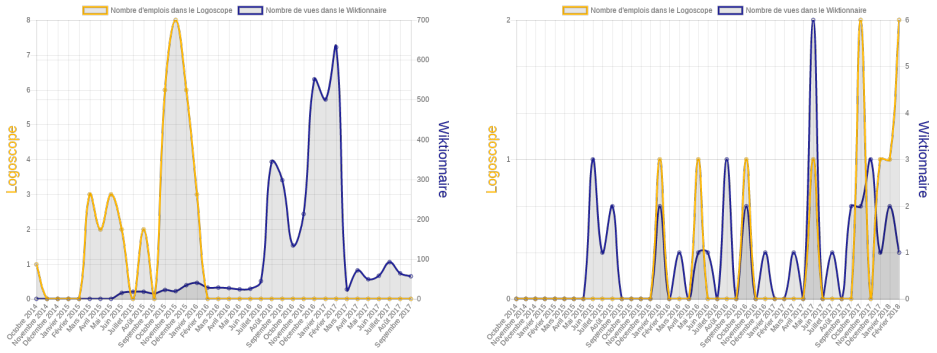
Except for the domain of sociolinguistics, psycholinguistics and discourse studies (e.g. Dufour and Rosier (2012)), the approaches adopted by linguists often implement a "language for language's sake" perspective, which prevents taking into consideration extra-linguistic realities. Concerning new words, this word-only oriented perspective is evident in a large scale of studies, ranging from the design of language specific rules (e.g. Corbin (1992)) to the description of textual phenomena (Gérard (2011, 2014)). This approach is certainly justified by the fact that most of the many new words created each year originate in a rhetorical intention (to denounce, to joke, to emphasise, etc.) rather than in a need to communicate about an extra-linguistic reality (technical, political, economical, environmental, etc.). The consequence is however that these extra-linguistic realities are not observed and studied.

With respect to such extra-linguistic realities new words can be classified in two groups according to their "cultural spread".

[38] This is about the "Charlie Hebdo" newspaper

Language	Non-cultural specific	Cultural specific
English	vlogging, retweet, brexit	Mx., beer o'clock
Italian	svapare, macaron, jihadista	enopirateria, sciarpata
Catalan	bitcoin, crowdfunding, wok	iaioflauta, xirucaire
French	téléverser, Zika, nomophobie	panthéonisable, CRDS

Table 7. Examples of cultural specific and non-cultural specific new words added in dictionaries between 2014 and 2016.



(a) *bitcoin* **(b)** *ethers*

Figure 11. Occurences of new words *bitcoin* and *ethers* in the *Logoscope* vs. lookups in the Wiktionary.

1. Words which have been created within a specific cultural area. They reflect realities which, at the time we experience them, are only happening in this area (*i.e.* realia). We assume therefore that they do not concern any other cultural area in the world.

2. Words refering to new realities which, mostly because of globalisation, also appear in other cultural areas.

Table 7 shows some examples from different languages for these two groups of new words. Since its launch in October 2014, the *Logoscope* is spotting new words designating emerging cultural specific realities. For example, in this period of time, a notable new word showing increasing frequency during several months was *zadiste*.

Zadiste is derived from "ZAD" which is an acronym for "zone à défendre" (area to be protected). The *Logoscope* data showed that after its appearance its frequency in the newspaper articles was relatively stable during several months. The thematic analysis revealed that it mostly appeared in textual contexts related to social conflicts. Thus the tool not only documented the appearance of this new word but also showed how its propagation went along with the establishment and consolidation of a new social conflict.

Other examples of how lexical creation testifies of political events are the new words *dealmaker* which was first detected in November 2016 when Donald Trump was elected president of the United States, and *Bruxellistan* which appeared between January and March 2016 when a suspect from the Paris terror attacs was arrested several months later in Brussels. Here the word *Bruxellistan* arguably also shows a rhetorical intention, namely to denounce a situation in Brussels and to put it on a level with a similar situation in London (which at some point was called *Londonistan*).

A word creation accompanying an economic event is *bitcoin*. Figure 11 shows when it was observed by the *Logoscope* (yellow lines) compared to when it was looked up in the Wiktionary (blue lines).[39] According to this it first appeared in the (monitored) press before it was available in the Wiktionary. In contrast, *ethers*, a word creation related to the same economic event was visible in the Wiktionary first.

We also found that word creations ending in *-gate* were reliable indicators for various scandals: *volkswagengate* and *dieselgate* accompanied the diesel emission scandals and *fillongate* appeared when a French presidential candidate was suspected of corruption.

[39] `https://fr.wiktionary.org/`

Another interesting new word detected was *Hauts-de-France* which designates the name of a newly created administrative unit (a region) in France. Monitoring this word with the *Logoscope* allowed to testify of continuity and disruptions in this process of geopolitical change. Another administrative unit created at this point in France was the region which later was called *Grand-Est*. Whereas obviously *Hauts-de-France* was used by journalists, *Grand-Est* did not appear in the *Logoscope* collection, suggesting that this territory was less perceived as a genuine traditional or historical entity than the *Hauts-de-France*. Instead, we found several other denominations as for *e.g. alsace-champagne-ardenne-lorraine*, *alsace-lorraine-champagne-ardenne* or *alsace-champagne-ardenne*, which are all compositions made up of the names of the original regions.

Other word creations, like *bio-sourcé, bio-morale, bio-éthiciens*, are built using a particular prefix (in this case *bio* (organic)) which largely determines the new word's meaning. In the case of *bio* it also arguably indicates the development of a new social identity centered around an organic style of living.

Some other word creations arguably point to phenomena related to the perception of gender identity. Thus, many new words designating occupations (or occupational titles) appear with either a female morphology only or both female and male morphology. Some examples are *clippeuse, proffesseuse, startupeuse, youtoubeuse*.

These examples show that this data and way of monitoring lexical change may be of direct interest not only for lexicographers and linguists but also for journalists, historians and researchers in social and political sciences.

7 Conclusion

This paper described the *Logoscope* framework – a tool for the detection and documentation of French new words in online journalistic publications.

Compared to other available new word detection systems, the *Logoscope* belongs to a more elaborate category of frameworks consisting of a detection phase based on a methodologically well defined dynamic corpus and a documentation phase where the detected and documented new words are added to a linguistic knowledge base. In this paper we described both the corpus and methods used for the detection and the theoretic linguistic and extra-linguistic principles guiding the documentation of the new words.

We showed how the *Logoscope* benefits from recent developments of natural language processing techniques to largely automate both the detection and the documentation steps and to thereby allow a more extensive and comprehensive empirical study of the emergence of new words.

The *Logoscope* is targeted not only at the traditional user categories of such a tool (*e.g.* linguists and lexicographers) but more generally at user groups concerned with how various cultural, societal or political developments are reflected in journalistic publications (*e.g.* journalists but also social scientists or economists). We showed how this viewpoint determined the documentation of new words in the *Logoscope* framework. One consequence was our choice to deliberately restrict the description of the lexical features such that the annotation effort is reduced but a subsequent in depth linguistic or lexicographic analysis remains possible. In contrast, we opted for more extensive information about contextual features of the new words, as for example the topics addressed by the articles containing them. To our knowledge the *Logoscope* is the only neologism detection system providing this feature automatically.

The *Logoscope* new words base is publicly available and can be queried online at the address `http://logoscope.unistra.fr`.

References

Aho AV, Corasick MJ (1975) Efficient string matching: an aid to bibliographic search. Commun ACM 18(6):333–340

Beust P (2002) Un outil de coloriage de corpus pour la représentation de thèmes. In: JADT 2002 : 6emes Journées internationales d'Analyse statistique des Données Textuelles, France, pp 161–172

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

Borde D (2016) Tirons la langue : Plaidoyer contre le sexisme dans la langue française. UTOPIA, Paris

Boulanger JC (2010) Sur l'existence des concepts de "néologie" et de "néologisme". Propos sur un paradoxe lexical et historique. In: Cabré T, Domènech O, Estopà R, Freixa J, Lorente M (eds) Actes del I Congrés Internacional de Neologia de les Llengües Romàniques, Institut Universitari de Lingüística Aplicada, Barcelona, pp 31–74, URL http://www.academia.edu/download/31304144/Actes_del_I_CongreIs_Internacional_de_Neologia_de_les_LlenguI%CB%86es_RomaIniques.pdf#page=31

Boussidan A, Ploux S (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 MT, Domènech M, Estorpà R, Freixa J, Solé E (2003) L'observatoire de néologie: conception, méthodologie, résultats et nouveaux travaux. L'innovation lexicale pp 125–147

Cabré T (2000) La neologia com a mesura de la vitalitat interna de les llengües. In: Cabré T, Freixa J, Solé E (eds) La neologia en el tombant de segle. Barcelona: Universitat Pompeu Fabra, Biblioteca de Catalunya, Barcelona, pp 85–108

Cabré T (2015) La neologia: un nou camp a la cerca de la seva consolidació científica. Caplletra Revista Internacional de Filologia (59):125–136

Cabré T (2016) Per què és relativament fàcil de detectar neologismes i tan complicat de definir què són: breu apunt epistemològic. Mots d'avui, mot de demà, Observatori de Neologia, Institut Universitari de Lingüística Aplicada, Barcelona, pp 127–132

Calvet LJ (2016) La Méditerranée: mer de nos langues. Langage et société 158(4):328

Cartier E (2011) Néologie et description linguistique pour le TAL. Langages 183:105–117

Chang CC, Lin CJ (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

Corbin D (1992) Sens et Définition: De la Compositionnalité du Sens des Mots Construits (Réponse a Claire Vanderhoeft). Lingvisticæ Investigationes 16(1):189–218, DOI http://dx.doi.org/10.1075/li.16.1.10cor, URL http://www.jbe-platform.com/content/journals/10.1075/li.16.1.10cor

Dal G, Namer F (2016) À propos des occasionnalismes. In: 5e Congrès Mondial de Linguistique Française - CMLF 2016

Dufour F, Rosier L (2012) Introduction. Héritages et reconfigurations conceptuelles de l'analyse du discours « à la française » : perte ou profit ? Langage et société 2(140):5–13

Estopà, R, Cabré Castellví, M-T (2004) Metodologia del treball en neologia: criteris, materials i processos

Falk I, Bernhard D, Gérard C (2014) From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. In: LREC - The 9th edition of the Language Resources and Evaluation Conference, Reykjavik, Iceland, Proceedings of the International Conference on Language Resources and Evaluation

Gérard C (2011) Lexical creation, sense and textuality: theories and analysis. PhiN (56):1–30

Gérard C (2014) Sémiotique interprétative des créations verbales. In: Driss Ablali DD Sémir Badir (ed) Documents, textes, oeuvres. Perspectives sémiotiques., Presses universitaires de Rennes

Gérard C, Falk I, Bernhard D (2014) Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique ? In: Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014), Berlin, Germany, SHS Web of Conferences., vol 8, pp 2627 – 2646, DOI 10.1051/shsconf/20140801208

Guilbert L (1975) Les travaux de linguistique en matière de néologie. In: Dupuis H (ed) L'aménagement de la néologie, actes du Colloque international de terminologie, L'éditeur officiel du Québec, Québec, pp 121–131

Gérard C, Lacoste C (2016) « Mots de la Grande Guerre » : créations inaperçues et usages réels dans les écrits de combattants. Étude de lexicologie textuelle. La première guerre mondiale et la langue

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explorations 11(1):10–18

Hoffman MD, Blei DM, Bach FR (2010) Online Learning for Latent Dirichlet Allocation. In: NIPS, pp 856–864

Hoffman MD, Blei DM, Wang C, Paisley J (2013) Stochastic variational inference. The Journal of Machine Learning Research 14(1):1303–1347

Humbley J, Sablayrolles JF (2016) Néologie et corpus. No. 10 in Neologica. Revue internationale de néologie

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

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

Lemnitzer L (2010) Neologismenlexikographie und das internet. Lexicographica 26:65–78

Lemnitzer L (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

Medelyan O, Witten IH (2008) Domain-independent automatic keyphrase indexing with small training sets. J Am Soc Inf Sci Technol 59(7):1026–1040

Mejri S, Sablayrolles JF (2011) Présentation: Néologie, nouveaux modèles théoriques et NTIC. Langages (183):3–9

Neveu F (2008) Pour une description terminographique des sciences du langage. Cahiers du CIEL 2008:87–104

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

Peschel C (2002) Zum Zusammenhang von Wortneubildung und Textkonstitution. Niemeyer

Pruvost J, Sablayrolles JF (2012) Les néologismes, 2nd edn. PRESSES UNIVERSITAIRES DE FRANCE - PUF

Reutenauer C (2012) Vers un traitement automatique de la néosémie : approche textuelle et statistique. PhD thesis, Université de Lorraine

Rey A (1976) Néologisme, un pseudo-concept ? Cahiers de Lexicologie 28(1):3–17

Roche S, Bowker L (1999) Cenit : Système de détection semi-automatique des néologismes. Terminologies nouvelles (20):12–16

Romary L, Salmon-Alt S, Francopoulo G (2004) Standards going concrete: from LMF to Morphalou. In: Workshop Enhancing and Using Electronic Dictionaries, Geneva, Switzerland

Sablayrolles JF (2006) Terminologie de la néologie: lacunes, flottements et trop pleins. Syntaxe et sémantique (7):79–89

Sablayrolles JF (2010) Neologia : un dictionnaire néologique sous forme de base de données. Instituto de Letras da Universidade federal da Bahia, pp 221–235

Sablayrolles JF (2011) Neologia : un dictionnaire néologique sous forme de base de données. Os dicionarios, fontes, métodos et novas tecnologias pp 221–235

Siebold O (2000) Wort-Genre-Text: Wortneubildungen in der Science Fiction. Gunter Narr Verlag

Solé E (2002) Textos i neologismes. In: Cabré, M Teresa and Freixa, Judit and Solé, Elisabet (ed) Lèxic i neologia, Observatori de Neologia. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra, Barcelona, pp 77–88

Steyvers M, Griffiths T (2007) Probabilistic Topic Models, Lawrence Erlbaum Associates

Swiggers P (2010) Terminologie, terminographie et métalangage linguistiques: Quelques réflexions et propositions. Revue roumaine de linguistique 55(3):209–222

Wulf Oesterreicher, Peter Koch (2001) Gesprochene Sprache und geschriebene Sprache/Langage parlé et langage écrit. Lexikon der Romanistischen Linguistik pp 601–604

Řehůřek R, Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, University of Malta, Valletta, Malta, pp 46–50