Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2117–2125
Marseille, 20-25 June 2022

A Methodology for Building a Diachronic Dataset of Semantic Shifts and its Application to QC-FR-Diac-V1.0, a Free Reference for French

David Kletz†, Philippe Langlais†, François Lareau‡, Patrick Drouin‡
Université de Montréal
†Department of Computer Science and Operational Research (DIRO) / RALI
‡Department of Linguistics and Translation
{david.kletz, philippe.langlais, francois.lareau, patrick.drouin}@umontreal.ca

Abstract

Different algorithms have been proposed to detect semantic shifts (changes in a word meaning over time) in a diachronic corpus. Yet, and somehow surprisingly, no reference corpus has been designed so far to evaluate them, leaving researchers to fallback to troublesome evaluation strategies. In this work, we introduce a methodology for the construction of a reference dataset for the evaluation of semantic shift detection, that is, a list of words where we know for sure whether they present a word meaning change over a period of interest. We leverage a state-of-the-art word-sense disambiguation model to associate a date of first appearance to all the senses of a word. Significant changes in sense distributions as well as clear stability are detected and the resulting words are inspected by experts using a dedicated interface before populating a reference dataset. As a proof of concept, we apply this methodology to a corpus of newspapers from Quebec covering the whole 20th century. We manually verified a subset of candidates, leading to QC-FR-Diac-V1.0, a corpus of 151 words allowing one to evaluate the identification of semantic shifts in French between 1910 and 1990.

Keywords: Semantic Shift Identification, Semantic Neology, Neology, Reference Corpus, Evaluation

1. Introduction

The meaning of a word is subject to variation over time. Many phenomena can trigger these changes, such as the appearance of new technology or social changes. Thus, an existing word can be assigned a new meaning, have a meaning withdrawn or remain stable between two given dates. The study of semantic shift is a field that focuses on these changes in meaning over time. Semantic shift, or semantic neology, is different from lexical neology. Detecting lexical neologisms across periods of a diachronic corpus boils down to detecting new, recurrent tokens in the latest period, which is a fairly straightforward process. With semantic neology however, the same token appears with different meanings in different periods. Detecting these cases is more involved. In this paper, we use the term *semantic shift* to refer to different cases, as explained below.

The first case is the appearance of a new word sense. This can be further divided into new polysemy and new homonymy, depending on how the new sense is related to an existing one. But the difference between the two is basically one of degree, and there are cases that sit on the fence. For example, consider the French noun *souris* ('mouse') between 1900 and today. The first sense for this word is defined as[1] *petit mammifère rongeur* ('small rodent mammal'). That meaning existed in 1900 and is still present today. A second meaning is *périphérique d'entrée relié à l'ordinateur* ('input device connected to the computer'). This meaning appeared at some point during the 20th century and was unknown in 1900. Whether this is case of polysemy or homonymy is debatable: the two meanings are seemingly unrelated and appear in completely different contexts, but there is still a metaphorical relation between the two, based on a vague resemblance between the peripheral and the animal. However, this distinction is not relevant for our purposes.

The second case is a shift in the dominant meaning of a word in the usage between two periods. For example, the Quebec French word *foulard* ('scarf') can denote a cloth worn around the neck to protect against the cold or a cloth worn on the head by Muslim women. Both senses existed in French in the 1970s and the 1990s, but we observed a sharp shift in their relative usage frequency between these two periods.

The third case is when a word becomes part of a new idiom. For example, the French adjective *froid* ('cold') became part of the new idiom *Guerre froide* ('Cold War') in the last half of the 20th Century.

One aim of diachronic studies is to track such changes over time.

Different algorithms have been proposed to detect semantic shifts in a diachronic corpus (see section 2), but surprisingly, no reference corpus has been designed so far to evaluate them, leaving researchers to fallback to troublesome evaluation strategies. In particular, Hamilton et al. (2016) establish a list of 28 word-forms (the largest reference set we found in our literature survey) known to have undergone a clear-cut change in meaning according to various information sources. While this allows to check the presence of those reference words in the list of words flagged by the algorithm as having evolved in meaning over time (therefore serv-

[1] All French lexicographic definitions mentioned in this paper are from Trésor de la Langue Française Informatisé (ATILF/CNRS–Université de Lorraine, accessed March 2021), `http://atilf.atilf.fr`. Some have been simplified for readability.

ing as a proxy for recall), one cannot tell if the other words retrieved have actually changed in meaning (i.e., measure of precision).

In this paper, we introduce a methodology for the semi-automatic construction of a reference dataset, which contains words that have undergone changes in meaning over time, and words which have not. In a nutshell, we rely on a predefined list of senses for words and a state-of-the-art word-sense disambiguation algorithm to label each occurrence of a word form in a diachronic corpus. Significant changes in sense distributions as well as clear stability are detected. The results are then validated by experts using a dedicated interface before populating a reference dataset.

As a proof of concept, we applied our methodology to a diachronic corpus of newspapers from Quebec covering the whole 20th century, leading to QC-FR-Diac-V1.0, a freely available resource[2] that contains 151 words allowing one to evaluate the identification of semantic shifts in Quebec French between 1910–20 and 1990–00.

This article is organized as follows. In section 2 we describe related work. We present our methodology in section 3 and describe our word-disambiguation model for French in section 4. We report the deployment of our methodology in section 5, leading to QC-FR-Diac-V1.0, a resource we analyze in section 6. We conclude our work in section 7 where we discuss future works.

2. Related works

Studies in diachronic semantics aim to propose an analysis of the meaning of a word between two selected periods of a so-called diachronic corpus.

Among the most used diachronic corpora for such tasks are the Brown corpus (Francis and Kucera, 1979), notably used by Mair and Leech (2006), the Corpus of Historical American English (COHA) (Davies, 2015), used for instance by Hamilton et al. (2016), and the Google N-Grams corpus (Lin et al., 2012), used by Hellrich and Hahn (2017).

Early on, Renouf (1993) proposed to exploit the frequency variation of associations to identify semantic shifts. This technique has been used under various forms and refined over the years (Kilgarriff et al., 2004; Castellvi et al., 2011; Cartier, 2009) but slowly migrated towards vector-based approaches (Turney and Pantel, 2010). The advent of embeddings exploiting deep learning techniques has lead to significant advances in the last few years. Recent algorithms have been proposed for identifying changes in word senses over time (Hamilton et al., 2016; Ryskina et al., 2020). In a nutshell, a diachronic corpus is sliced into periods (or epochs) of non-overlapping texts (a period of 10 years is often used), on which static-word embeddings (Mikolov et al., 2013) are computed. Significant changes in the neighbourhood of token forms are then used as the main signal to detect changes in meaning.

The validation step in semantic shift studies is challenging and is typically based on lexicographic definitions from common dictionaries. As a canonical illustration, Hamilton et al. (2016) rely on a list of 28 words whose change in meaning are known. Those are words used in previous studies on semantic shift, as well as words from the Oxford English Dictionary (Simpson et al., 1989) where the definition is explicitly labelled as “obsolete”, indicating that a sense has gone out of use. In addition to those words, they use a list of 10 words that have undergone the most significant changes in meaning according to their methodology, and they refer to the Oxford English Dictionary to confirm or refute the proposed change in meaning. Noticing the lack of resources for evaluating computational models of word sense change over time, Schlechtweg et al. (2018) propose a framework to annotate diachronic changes in word meaning. They put this framework at use to produce a dataset of 22 German words called DURel. The words in DURel were originally selected from a list of words known to have undergone a semantic shift over time across 2 two time periods (1750–1799 and 1850–1899). Schlechtweg et al. (2019) rely on this dataset as well as SURel (Hätty et al., 2019), a list of 22 German words that have undergone a semantic shift between domains (and not over time), in order to propose a comparison of the models of diachronic lexical semantic changes detection

It is questionable that firm conclusions regarding algorithms can be drawn based on such narrow references, which motivated our study in the first place.

3. Methodology

Initially, our plan was to use existing dictionaries to identify words whose definitions are dated, allowing us to retrieve all words whose meaning has changed over time. Rapidly, we realized that this would be no picnic. It is actually not easy to access programmatically many of available dictionaries, not to mention the legal issues in doing so. In addition, electronic dictionaries that list obsolete meanings of a word are not abundant. We considered the French Wiktionary [3] but soon realized that the entries are not easy to parse for our purpose and that very few definitions are actually dated, making it hard to use.

We then considered lexicographic projects dedicated to aligning different versions of dictionaries. Notably, we found the *Club d’orthographe de Grenoble*[4], which provides alignments between different versions of *Petit*

[2] Freely available at https://github.com/davkletz/QC-FR-Diac

[3] https://fr.wiktionary.org/wiki/Wiktionnaire:Page_d\OT1\textquoterightaccueil

[4] https://orthogrenoble.net/mots-nouveaux-dictionnaires

Larousse[5] and *Petit Robert*,[6] two popular French dictionaries. Unfortunately, this project is evolving slowly and only a few resources of interest to us are currently available.

There is also the *Nénufar* project (Bohbot et al., 2019), which aims to provide access to different versions of *Petit Larousse Illustré*. Unfortunately, the associated platform[7] gives access to a SPARQL engine that does not seem functional yet. On top of that, the project in its current state could not help us list enough useful information for our needs.

We therefore turned to a semi-automatic approach to list words that have evolved over time. It relies on the availability of two resources: a diachronic corpus, i.e., a collection of texts produced over a period of interest, and a word-sense disambiguation (WSD) technique that labels each word occurrence with a sense tag from a predefined set of senses. Such resources exist for several languages, including French.

We first apply a WSD model to the full diachronic corpus. This allows to collect, for each time period, the distribution of senses for each word. We then search for significant changes – or stability – in the distribution of senses for each word. The resulting words are then checked manually to validate whether their senses have actually evolved over time. Since this is not an easy task, we developed a dedicated interface to make the annotation process easier.

4. WSD for French

Our methodology relies on the availability of a WSD engine, but most works on WSD have been developed for English. We are not aware of a WSD system readily available for French, the language we are interested in here.

4.1. Main model

Fortunately, Vial et al. (2019) proposed a disambiguation pipeline easy to reproduce which relies on WordNet (Fellbaum, 1998) to determine the possible meanings of a word and on a BERT architecture (Devlin et al., 2019) to conduct the disambiguation.

A few pre-trained BERT models are available for French. For our study, we compared FlauBERT (Le et al., 2020) to CamemBERT (Martin et al., 2020), both available from the *Hugginface* library (Wolf et al., 2020).[8] We fine-tuned those models on the WSD task of the FLUE data collection (Le et al., 2020) for 20 epochs. Each fine-tuning stage took roughly a day (25 hours) on a GPU-equipped desktop computer. We tested the performance of both models on the French SemEval 2013 task 12 corpus (Navigli et al., 2013). Results are reported in the first column of Table 1 and one can see that CamemBERT has a small advantage. For French, Vial et al. (2019) reported an accuracy of 53.53% for their FlauBERT-large configuration, and 52.06% with a camemBERT-small configuration. Our performances are slightly off the performances reported with their best configuration, likely because we did not pay much attention to the meta-parameters.

However, it is noteworthy that performances are much lower than those typically reported for transformer-based solutions in English. For instance, Vial et al. (2019) report over 79.0% of accuracy on the concatenation of SE2, SE3, SE07 17, SE13 and SE15 WSD tasks of the evaluation campaigns Senseval/SemEval. This suggests that deploying our methodology for English should be easier, something we leave for future investigations.

4.2. Filtering for better precision

Since our goal is not WSD in itself, but rather to associate accurately a sense to some of the tokens available, we can afford to filter words to increase accuracy. We did this by imposing a threshold (0.5) on the best prediction scored by the softmax function for a given token. While there are other ways of improving accuracy, we observe in the second column of Table 1, that this increases the precision of the models to up to 70% while silencing roughly 40% of the predictions made, a satisfactory compromise for our needs. The best results are again observed with the CamemBERT model, the one we finally deployed.

Model	**Acc.**	**Filtered Acc.**	*(% disamb)*
FlauBERT	52.4	66.8	*(66.2)*
CamemBERT	52.9	70.9	*(59.9)*

Table 1: Performance of two models (large versions) fine-tuned for 20 epochs on the WSD training material of the FLUE benchmark and tested on the French SemEval 2013 task 12 corpus. The first column shows complete accuracy (computed on the full test set), while the second column shows the accuracy measured on the only words where the prediction is scored at least 0.5 by the softmax function. Figures in parentheses indicate the percentage of decisions taken.

5. Genesis of QC-FR-Diac-V1.0

We describe the different steps we deployed to develop QC-FR-Diac-V1.0, our reference benchmark for semantic shift studies in French.

5.1. Diachronic corpus

The corpus we used in this study comprises 196 sources of publications from Quebec accumulated over a pe-

[5] https://www.larousse.fr/dictionnaires/francais

[6] https://www.lerobert.com

[7] http://nenufar.huma-num.fr/avancement/

[8] We used the large version of each model: a 24-layer configuration with 16 attention heads, and an hidden size of 3072.

	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–00
# Tokens (10^6)	132.1	139.1	162.4	174.3	177.1	218.2	242.6	303.6	248.3
# Types (10^6)	3.9	5.7	4.9	4.9	11.2	7.0	3.9	5.0	3.2
# Hapax (10^6)	3.2	4.8	4.0	4.0	9.5	5.7	2.9	3.6	2.3
# Types w/freq $\geq$ 200 (10^3)	22.8	23.7	26.6	27.8	30.8	34.2	35.1	45.1	36.7

Table 2: Main characteristics of the diachronic corpus *Le Devoir* used in this study. Nine non overlapping slices of 10 years have been considered over the period 1910–2000.

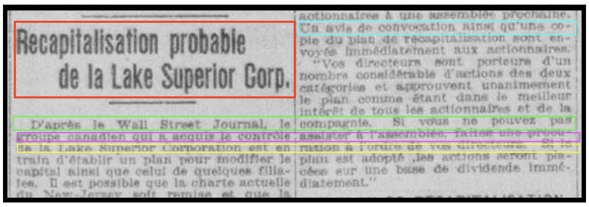
riod starting in 1800 and ending in 2000. It is made available by the CO.SHS project, an open infrastructure for the humanities and social sciences.[9] Most of the sources are journalistic, but not all of them. Some are published daily, others weekly or monthly, and the topics vary widely (religion, cuisine, local news, general news, people, etc.). We only considered French publications here.

All documents have been digitized and we used an already available textual output produced by a professional Optical Character Recognition (OCR) system.[10] The quality of the OCR varies a lot depending among other things on the quality of the original document. In particular, we noticed problems with old, yellowed and sometimes damaged papers, with multi-columns publications, or with documents that featured frequent font changes. Such problems are illustrated in Figure 1.

OCR accuracy has significantly improved recently, mainly due to the blooming of deep learning models (see for instance (Li et al., 2021) for a recent pre-trained transformer solution), but deploying such models on our large data collection would incur excessive efforts far beyond the scope of this study. The presence of OCR problems justifies the filtering strategy we have applied, coupled to a manual validation of candidate words.

We needed texts evenly distributed over a long period of time, and of homogeneous quality. Therefore, we considered daily newspapers in the first place. Among all the sources, only six were dailies covering a long enough period (over 50 years), of which only one, *Le Devoir*,[11] presented a regular number of tokens over time periods. Therefore, we considered this single source, which covers the years 1910 to 2000.

We divided the data into 9 non-overlapping time periods of 10 years. The main characteristics of this resource are reported in Table 2. We observe a rather large number of token types per period (around 4 millions), in part due to the nature of the corpus (daily news with a lot of proper nouns), but mainly because of the numerous OCR errors, which is corroborated by the very high ratio of hapax types over tokens.



Recapitalisation probable de la Lake Superior Corp.

D'après le Wall Street Journal, le groupe canadien qui a acquis le contrôle de la Lake Superior Corporation est en train d'établir un plan pour modifier le capital ainsi que celui de quelques filiales. Il est possible que la charte actuelle du New-Jersey soit remise et que la

actionnaires à une assemblée prochaine. Un avis de convocation ainsi qu'une copie du plan de recapitalisation sont envoyés immédiatement aux actionnaires. "Vos directeurs sont porteurs d'un nombre considérable d'actions des deux catégories et approuvent unanimement le plan comme étant dans le meilleur intérêt de tous les actionnaires et de la compagnie. Si vous ne pouvez pas assister à l'assemblée, faites une procuration à l'ordre de vos directeurs. Si le plan est adopté ,les actions seront placées sur une base de dividende immédiatement."

```
RBcapitallsatlon probable de la Lake Superior Corp.
avait ivnîreprinee aux usines de bt, Ca- j faveur d#5 pori-llr.-» d'actioru? ordirlftlr#» therina»;
il rn» fait plaisir de air# que 1 actueBé», soit au taux d'un* pour une. Ofis ligne», censista.nl
surtout d# la- ;  ^ tmU,^ s<*t 86,996 aettan» ''A" «t
biew'.x d# command# et inMaflatlon gé« j io,606 «otioni» "B", demeurera dans le nfirai#
électrique, dont le marolifi s é-I trÀw .pour le» besoin» fulUîi» de la oom-ter:d
oontlnuellement au Panada, ont
été tri>« favoraidement accueülies pafi; si le plan est adopté, les dfirecte<9c>-» le» ollent». î
ont Vin trillion de reoommandvr le pale-
<93>Vert directeur» fittidH-nt depuis quel- | nw-r.t d'«n dividende de *3 ipar action . qua
iemt» un plan pour modifier !»i -A" et espèrent pouvoir «n dont muer raht-gfinfiraî, en sera
ie pnésrident eijpn»! d< la oompagnl#: ce plan est, paiemaxit.
gérant-général, iinair, tenant terrotné et Son-, soumis aux!  ...................- ¦¦      ....
actionnaire» S une aasemblfi» prochaine. I ~r"
ÏTn avl» de convocation ainsi qu'un# cd- ; pie du plan de rficapitAlieatk>n sont en- ; vpyéé'
Immédiatement aux actionnaire» !
¦<95>Vos directeurs sont porteur» d'un! nombre considérable (l'action* de» deux '
catégories et approuvent unanimement I le plan comme étant dan* le meilleur [ intfirft' rie
tous le» actionnaire» et de la D'après le Wall Street Journal, le cobipsurnle. Si vtws ne
pouvez P»'' groupe canadien qui a acquis la contrôle iasolaier ft rassemblée, fafte» une
propu-de la Lake Superior CoritoraÉlon est en [ ration A l'ordre de vos directeurs, b, le
train d'établir un plan pour modifier leiplan est adopté .le»
action» «fWit plft-i-apltai iilnst que celui de quelque» ftîlft-
|eéev* sur une btu» de dividende Immé-le«. U «si iKisalbl» que la
chart# actuelle I dlatement."
du New-Jersey soir remis# et que l
```

Figure 1: Example of a publication (top) poorly handled by OCR (bottom). Coloured frames are manually added here for visual alignment.

5.2. WSD Deployment

While our ultimate objective is to date diachronic phenomena precisely, for this paper, we only looked whether we could trace a change in meaning for tokens between time periods 1910–20 and 1990–00, arguably an easier setting since those periods are distant in time. Comparing adjacent periods would lead to a more precise technique of dating of the appearance (or disappearance) of senses. This is left as future work.

We are not interested in words that appeared (or disappeared) over time and we thus focus on words present in the two time periods of interest. In order to reduce the impact of disambiguation errors, and to ensure a strong signal, we further impose that each type considered receives at least 200[12] disambiguations per time period studied. Table 3 shows the statistics of

[9] https://co-shs.ca

[10] Several OCR algorithms may have been used for different publication sources.

[11] https://www.ledevoir.com

[12] By choosing such a threshold, we ensure that there is a 90% chance that at least 65% of the occurrences have been well disambiguated.

the tokens and types disambiguated. Up to 22% of the types were disambiguated per period, which represents a large enough portion of types for our goal.

	Tokens		Types	
Period	**Total**	**Disamb.**	**Total**	**Disamb.**
1910–20	97.9	27.9 (28%)	3.6	0.80 (22%)
1990–00	193.2	51.5 (27%)	3.3	0.73 (22%)

Table 3: Quantification of the number of disambiguations in terms of tokens and types. Figures are in millions. The tokenization is performed by the WSD pipeline (section 4), which slightly differs from the one used for compiling statistics in Table 2.

5.3. Selection of candidates

We are interested in identifying words which show changes in meaning, as well as some that do not. This is done by comparing the distribution of senses of a word of interest.

First, we identified 6640 words that received at least 200 disambiguations in both time periods 1910–20 and 1990–00. To avoid spurious senses (labelling errors are frequent, especially due to our noisy input), we only consider senses that represent at least 10% of the occurrences of a given word in a given time period. Among these 6640 words, 5356 had at least one sense representing 10% or more of the occurrences. From these words, we identify those that are stable between the two time periods and those that have changed in meaning.

First, we address words that have undergone a change in meaning. As an illustration, Table 4 compares the distribution of three meanings of the word *émissions* ('emissions') over the two time periods. The first meaning was present in 92% of the occurrences we could label in the first time period, but only 10% in the second one. Hence, we consider that this meaning has changed over time.

Sense	**1910–20**	**1990–00**
'issue'	92%	10%
'broadcast'	<10%	42%
'television program'	<10%	37%

Table 4: Sense distribution of the word type *émissions* over the two studied time periods. The sense 'issue' refers to the acception 'to put into circulation' and is largely predominant in the earlier time period, while the two other senses were much less. Since the difference of appearance of this sense over the two time periods (92% − 10%) is larger than 70%, we consider *émissions* a candidate word with a semantic shift.

According to the accuracy of the model and the minimal number of occurrences (200), we reach at least 65% of good disambiguations. Thus, in order to overcome the errors of disambiguation, we consider that a sense appears (or disappears) if a difference in distribution of at least 70% for this sense is observed between the two time periods (this way we ensure that the appearance of this meaning is indeed due to its use in the corpus). There are 1521 words with at least one sense with a difference in distribution of at least 70% between 1910–20 and 1990–00.

Finally, among those, we consider only the ones that have at least one sense that underwent such a variation and were absent in 1910–20 (according to the previous definition). This yields a list of 119 words that we suspect to have undergone a change in meaning, including *émissions*.

Similarly, we search for stable candidates, i.e., words with no change in meaning. As an illustration, the word *excédent* ('surplus') keeps between 1910–20 and 1990–00 only one definition: *quantité de quelque chose qui dépasse une quantité donnée (longueur, volume); ce qui se trouve en surplus* ('quantity of something which exceeds a given quantity (length, volume)'). Therefore, we label it as stable between the time periods 1910–20 and 1990–00.

A word is considered stable if and only if no sense appeared or disappeared between 1910–20 and 1990–00. Among the 6640 words that received at least 200 disambiguations in both time periods, 4652 have no sense appearance or disappearance. Furthermore, we expect a stable word to keep a certain stability in sense usage, i.e., an absence of variation of more than 10% of all senses. Applying this filter, we end up with 3587 stable words with two or fewer meanings, of which 2463 are monosemic.

Table 6 illustrates words which have likely undergone a change in meaning over the two periods of interest, while Table 5 shows words that likely have not.

Token	**Meaning 1**	**Meaning 2**
quatre	'cardinal'	—
donc	'therefore'	'consequently'
symphonie	'symphony'	—
privilège	'privilege'	'prerogative'
couteau	'knife'	—
témoigner	'testify'	'express'

Table 5: Examples of semantically stable candidate words between time periods 1910–20 and 1990–00.

5.4. Human Validation

In order to check the quality of the data obtained, we performed a manual validation of the candidate words we obtained. The validation of a meaning being rather laborious, we decided to stick to the inspection of the :

- 119 words that likely underwent a change in meaning

Token	1910–20	1990–00	Domain
tissu	'material'	'framework (structure)'	industry
émissions	'putting into circulation'	'TV programs'	technology
direct	'direct'	'performed live'	technology
union	'combination'	'USSR'	geopolitics
nations	'states'	'UN'	geopolitics

Table 6: Candidate words that underwent a semantic shift between 1910–20 and 1990–00, with their meaning in both time periods. The last column specifies the domain of the new sense.

- 150 words we randomly selected from the list of likely stable words, all monosemic (in order to limit the number of meanings to be validated to confirm stability)

leading to the human evaluation of a total of 269 words. Each example was presented to judges in the form of a word–meaning pair (noted respectively w and m), for which they had to answer the following question:

> May the word w have been used
> with the sense m in 1910–1920?

To answer this question, judges had access to the WordNet definition for the sense m, as well as contexts (one at a time) where token w is expected to have appeared with meaning m. Contexts gather 30 words before and after the candidate token and are extracted from the 1990–00 time period. Judges were allowed to skip a given example during the annotation.

6. Analysis

Two experts in linguistics, the two last co-authors of this paper, conducted the annotation. Obviously, objectivity and credibility would be enhanced by resorting to external evaluators, but this can be done in the future to refine the quality of this evaluation step.

The evaluators were presented with the same list of examples to annotate. As one can imagine, deciding whether a given sense can be associated to a word at a specific time period is not easy, and the judges flagged a fair number of examples as problematic. In fact, only 73% of the examples were judged non-problematic.[13] Details of the evaluation are in Table 7.

Somewhat as could be expected, the evaluators indicated they lacked confidence to decide a sense could not possibly have been used in 1910–20, thus encouraging them to click the problematic button. They also reported a tendency to feel that the senses proposed existed in that time period mainly because they exist today. They also noted that the sense tagging was often problematic, despite the efforts made to improve its quality. Indeed, many of the identified contexts did not correspond to correct sentences in French. In particular, contexts gathered from chunks of numerical entities interspersed with letters, which are typical of some OCR errors in our corpus, deceived the WSD model. Those cases were naturally selected as problematic.

The non-problematic examples that have been annotated by a single evaluator (28 + 44 = 72 for evaluator 1, and 39 + 9 = 48 for evaluator 2) were considered the ground truth, while for those for which both evaluators answered (36 + 54 = 90 examples) we kept only the ones where the decisions were identical. Among the 90 non-problematic examples annotated by both evaluators, 72 (80%) had the same annotation.

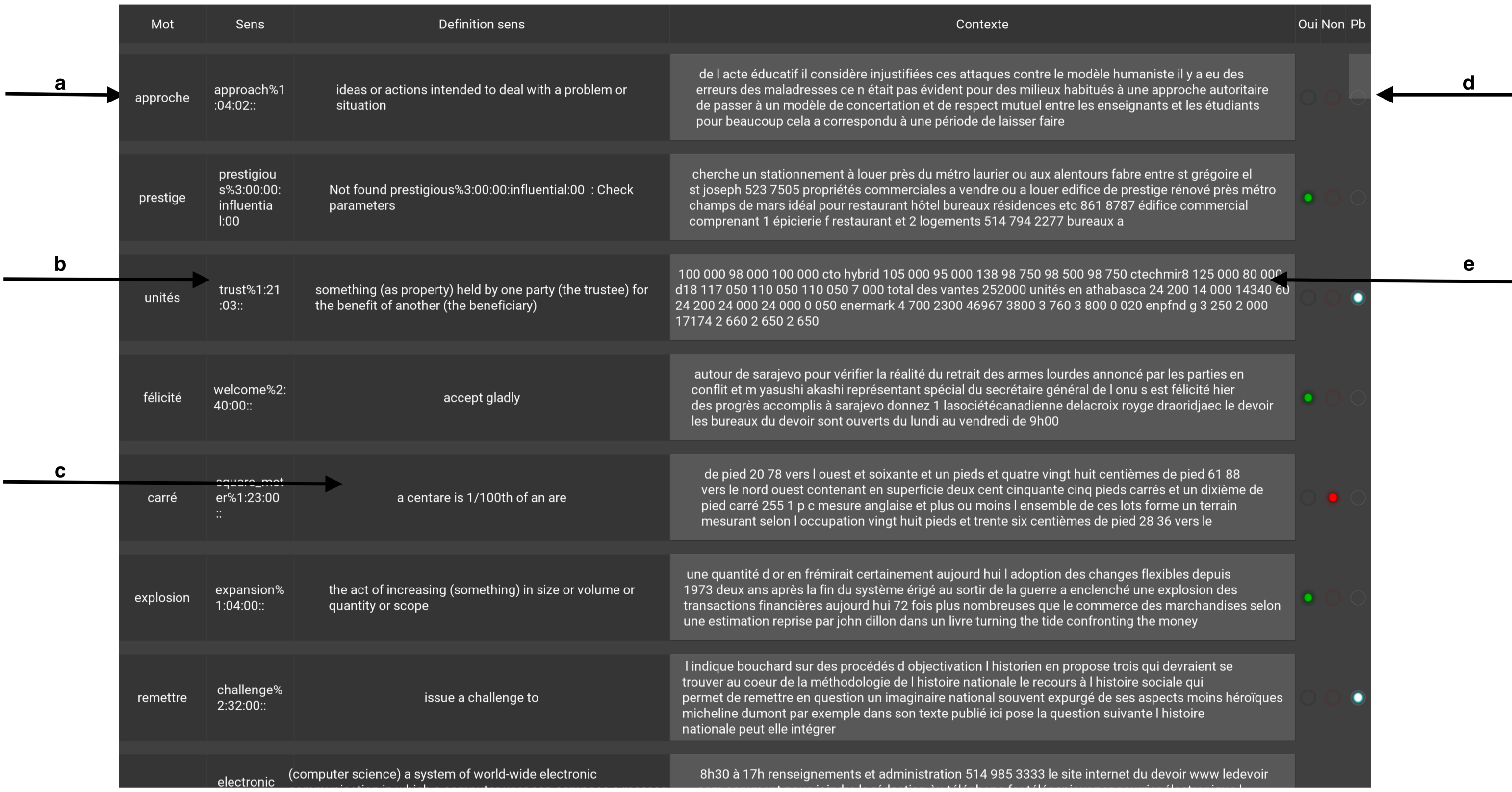
In the end, a total of 59 examples (45% of the non-problematic ones) have been validated as having undergone a change in meaning, and 92 examples (90% of the non-problematic examples) have been validated as stable. Those are the 151 words listed in QC-FR-Diac-V1.0.

Table 8 shows a selection of words presenting a semantic shift. Typically, such words are mainly concerned with specific domains, such as technology (16 examples), geopolitics (7 examples), and economy (9 examples). The 16 words we attributed to the technology domain are witnesses of major advances that occurred during the 20th century, with the arrival and democratization of radio and television (e.g., *station* in the sense of 'radio channel'), the development of analog and digital signal processing (*enregistrer* 'record', in

	Eval 1	Eval 2	1 & 2
Stable words			
examples	64	75	36
problematic	86	75	47
single annotation	28	39	
two annotations identical			33
agree w/ stability	27	34	32
Words with semantic change			
examples	98	63	54
problematic	21	56	12
single annotation	44	9	
two annotations identical			15
agree w/ change	43	1	15

Table 7: Details of the human evaluation: for example, the first line gives the details of the supposed stable words that have been annotated : the first evaluator annotated 64 of them, the second 75, and 36 were annotated by both the first and the second evaluators.

[13] This represents 82% of the examples predicted to have undergone a change, and 67% of the words predicted to be stable.



Mot	Sens	Definition sens	Contexte	Oui	Non	Pb
approche	approach%1:04:02::	ideas or actions intended to deal with a problem or situation	de l acte éducatif il considère injustifiées ces attaques contre le modèle humaniste il y a eu des erreurs des maladresses ce n était pas évident pour des milieux habitués à une approche autoritaire de passer à un modèle de concertation et de respect mutuel entre les enseignants et les étudiants pour beaucoup cela a correspondu à une période de laisser faire	○	○	○
prestige	prestigious%3:00:00:influential:00	Not found prestigious%3:00:00:influential:00 : Check parameters	cherche un stationnement à louer près du métro laurier ou aux alentours fabre entre st grégoire el st joseph 523 7505 propriétés commerciales a vendre ou a louer edifice de prestige rénové près métro champs de mars idéal pour restaurant hôtel bureaux résidences etc 861 8787 édifice commercial comprenant 1 épicierie f restaurant et 2 logements 514 794 2277 bureaux a	●	○	○
unités	trust%1:21:03::	something (as property) held by one party (the trustee) for the benefit of another (the beneficiary)	100 000 98 000 100 000 cto hybrid 105 000 95 000 138 98 750 98 500 98 750 ctechmir8 125 000 80 000 d18 117 050 110 050 110 050 7 000 total des vantes 252000 unités en athabasca 24 200 14 000 14340 60 24 200 24 000 24 000 0 050 enermark 4 700 2300 46967 3800 3 760 3 800 0 020 enpfnd g 3 250 2 000 17174 2 660 2 650 2 650	○	○	●
félicité	welcome%2:40:00::	accept gladly	autour de sarajevo pour vérifier la réalité du retrait des armes lourdes annoncé par les parties en conflit et m yasushi akashi représentant spécial du secrétaire général de l onu s est félicité hier des progrès accomplis à sarajevo donnez 1 lasociétécanadienne delacroix royge draoridjaec le devoir les bureaux du devoir sont ouverts du lundi au vendredi de 9h00	●	○	○
carré	square_meter%1:23:00::	a centare is 1/100th of an are	de pied 20 78 vers l ouest et soixante et un pieds et quatre vingt huit centièmes de pied 61 88 vers le nord ouest contenant en superficie deux cent cinquante cinq pieds carrés et un dixième de pied carré 255 1 p c mesure anglaise et plus ou moins l ensemble de ces lots forme un terrain mesurant selon l occupation vingt huit pieds et trente six centièmes de pied 28 36 vers le	○	●	○
explosion	expansion%1:04:00::	the act of increasing (something) in size or volume or quantity or scope	une quantité d or en frémirait certainement aujourd hui l adoption des changes flexibles depuis 1973 deux ans après la fin du système érigé au sortir de la guerre a enclenché une explosion des transactions financières aujourd hui 72 fois plus nombreuses que le commerce des marchandises selon une estimation reprise par john dillon dans un livre turning the tide confronting the money	●	○	○
remettre	challenge%2:32:00::	issue a challenge to	l indique bouchard sur des procédés d objectivation l historien en propose trois qui devraient se trouver au coeur de la méthodologie de l histoire nationale le recours à l histoire sociale qui permet de remettre en question un imaginaire national souvent expurgé de ses aspects moins héroïques micheline dumont par exemple dans son texte publié ici pose la question suivante l histoire nationale peut elle intégrer	○	○	●
	electronic	(computer science) a system of world-wide electronic	8h30 à 17h renseignements et administration 514 985 3333 le site internet du devoir www ledevoir			

Figure 2: Interface of annotation. One example is presented per line. (a) indicates the word w, (b) its meaning m (Wordnet synset), (c) provides the Wordnet definition of m, (d) are the buttons that the user might select (yes, no, problematic), and (e) shows one context (by clicking on it, the user can ask for a new context).

the sense of storing information in digital or analog format), or the development of the air travel industry (*vols* ‘flights’). Similarly, terms related to geopolitics are associated with new meanings linked to events such as the independence of states and their new names (*e.g.*, the name *Congo*, which was used to designate a river, now refers to two countries that obtained their independence in 1958 and 1960 respectively), or to major conflicts (*Froide* ‘Cold’, in reference to the cold war).

One novelty of QC-FR-Diac-V1.0 is that it lists words that did not change in meaning over the two periods of interest. We observed words from various parts of speech, including 15 verbs, 18 adjectives and 51 nouns (including 2 proper nouns). Here is a random selection of those words, along with their translation: *allemande* ‘german’, *persécutions* ‘persecution’, *bleue* ‘blue’, *fixant* ‘fixing’, *urgents* ‘urgent’, *conversation* ‘talking’, *musiciens* ‘musicians’, *Munich* ‘Munich’, *centre* ‘center’, *échantillons* ‘sample’.

We found it difficult to associate specific domains to those words, but to follow up on the domains we enumerated previously for which new meanings have been observed over time, we found a few examples of stable words in the same domains:

***allemande* ‘German’:** although Germany existed in various forms throughout the 20th century, the word for the associated nationality has remained stable.

***payant* ‘paid’:** word of the economic domain, always used in the context of monetization of services or goods.

***explosifs* ‘explosives’:** a technology that remained stable over the 20th century.

7. Conclusion

Diachronic studies are usually evaluated against words lists that are often very small. In this paper, we described a methodology which relies on a state-of-the-art WSD algorithm used to annotate tokens of interest

Word	Sense	Word	Sense	Word	Sense
stations	‘radio stations’	*manches*	‘tournament rounds’	*charme*	‘fashionable’
souverain	‘sovereign’	*émission*	‘television program’	*fournisseurs*	‘suppliers’
carré	‘square meter’	*armes*	‘nuclear weapons’	*consultations*	‘negotiation’
crédits	‘course credits’	*bande*	‘gaza strip’	*tournant*	‘turn of the century’
enregistrer	‘record’	*organisme*	‘agency’	*vols*	‘spaceflight’

Table 8: 15 of the 59 words that underwent a change in meaning between time periods 1910–1920 and 1990–00. The sense column indicates the sense that appeared for that word in the second time period.

in a given diachronic corpus into senses. The distribution of senses over each time period (10 years or so) are then compared to detect significant changes (or stability) over two time periods. We deployed our methodology on a diachronic corpus of daily news written in French and published in Quebec over the period 1910–2000. With carefully designed constraints, we identified a number of words that likely have undergone a change in meaning, as well as words that likely did not. Part of this material has been inspected manually, yielding QC-FR-Diac-V1.0 a freely available resource which contains 59 words with a change in meaning and 92 that we think are stable over the time periods 1910–20 and 1990–00. This is already larger than the lists of words currently being used for evaluation in works on dicahrony identification.

This work opens a number of interesting avenues. First, a more precise technique of dating of the appearance or disappearance of senses could be attempted by reducing the time distance between the time periods compared. Moreover, our data was mainly limited to the 20th century, but by relying on a corpus covering a longer period of time, one could expect to build a more substantial resource. Similarly, less stringent constraints could be tested in order to obtain a larger coverage. Especially, polysemous stable words should also be annotated and evaluated.

Also, our methodology could be deployed on any language for which WSD is mature enough. In particular, there should be no obstacles to deploying it for English, especially since the WSD technology for this language is of higher quality than the one we developed here for French.

Of course, the ultimate goal of our methodology is to obtain large enough reference lists to better evaluate diachronic detection algorithms, which remains to be done.

Acknowledgements

The study was conducted as part of project *NÉONUM: A first monitoring infrastructure for neology in Quebec French*, funded by the Social Sciences and Humanities Research Council of Canada (SSHRC).

Our corpus was kindly made available by the CO.SHS project, which is an open infrastructure for the humanities and social sciences. CO.SHS is financed by the Canada Foundation for Innovation as part of the Cyberinfrastructure initiative.

We would also like to thank the anonymous reviewers for their useful and constructive comments.

8. Bibliographical References

Bohbot, H., Frontini, F., Khan, F., Khemakhem, M., and Romary, L. (2019). Nénufar: Modelling a Diachronic Collection of Dictionary Editions as a Computational Lexical Resource. In *ELEX 2019: smart lexicography*, Sintra, Portugal, October.

Cartier, E. (2009). Utilisation des contextes dans le cadre dictionnairique : état des lieux, typologie des contextes, exemple des contextes définitoires. In *Actes des Huitièmes Journées scientifiques du Réseau de chercheurs Lexicologie, terminologie, traduction*, pages 619–632.

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

Davies, M. (2015). Corpus of Historical American English (COHA).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Francis, W. N. and Kucera, H. (1979). Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). ”diachronic word embeddings reveal statistical laws of semantic change”. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

Hätty, A., Schlechtweg, D., and Schulte im Walde, S. (2019). SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 1–8, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hellrich, J. and Hahn, U. (2017). Exploring diachronic lexical semantics with JeSemE. In *Proceedings of ACL 2017, System Demonstrations*, pages 31–36, Vancouver, Canada, July. Association for Computational Linguistics.

Kilgarriff, A., Rychlý, P., Smrz, P., and Tugwell, D. (2004). The sketch engine. In *Proceedings of the 11th EURALEX International Congress*, pages 105–115.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbe, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In *LREC*, Marseille, France.

Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2021). Trocr: Transformer-based optical character recognition with pre-trained models.

Lin, Y., Michel, J.-B., Aiden Lieberman, E., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea, July. Association for Computational Linguistics.

Mair, C. and Leech, G. (2006). Current change in english syntax. In Bas Aarts et al., editors, *The Handbook of English Linguistics*, pages 318–342. Blackwell.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de La Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States, July.

Mikolov, T., Chen, K., Corrado, C., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Navigli, R., Jurgens, D., and Vannella, D. (2013). SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Renouf, A. (1993). Sticking to the text: a corpus linguist's view of language. *Aslib Proceedings*, 45(5):131–136. Publisher: MCB UP Ltd.

Ryskina, M., Rabinovich, E., Berg-Kirkpatrick, T., Mortensen, D., and Tsvetkov, Y. (2020). Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 367–376. Association for Computational Linguistics.

Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, LA, USA.

Schlechtweg, D., Hätty, A., Del Tredici, M., and Schulte im Walde, S. (2019). A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy, July. Association for Computational Linguistics.

Simpson, J. A., Weiner, E. S., and al. (1989). The oxford english dictionary, volume 2.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Vial, L., Lecouteux, B., and Schwab, D. (2019). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Global Wordnet Conference*, Wroclaw, Poland.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

9. Language Resource References

ATILF/CNRS–Université de Lorraine. (accessed March 2021). *Trésor de la Langue Française informatisé (TLFi) [online resource]*. ATILF/CNRS.