Metalexicographical Investigations with the DiCo Database

Franck Sajous (CLLE, CNRS & Université Toulouse 2)
and Camille Martinez (Orthodidacte)

To cite this paper:

Franck Sajous and Camille Martinez (2022). Metalexicographical Investigations with the DiCo Database. *International Journal of Lexicography*, 35(1), pp. 75-106.

Metalexicographical Investigations with the DiCo Database

Abstract

This article presents DiCo, an inventory of the changes in the nomenclature of four French dictionaries (*Dictionnaire de l'Académie française*, *Dictionnaire Hachette*, *Le Petit Larousse* and *Le Petit Robert*). For each modification recorded, DiCo provides additional information on the microstructural level, such as the linguistic labels included in the article where the change occurred. Based on a manual comparison of the successive editions of a given dictionary, DiCo can be the starting point for quantitative and qualitative metalexicographical studies. The description of the diachronic evolution of a dictionary and the comparison of different dictionaries reveal that not only does lexicographical change reflect language evolution, but that the content of a dictionary is also bound to the editorial policy of a publishing house, itself subject to change. Given the scarcity of information on this topic provided to the general public by French publishing houses, a resource facilitating metalexicographical investigations is particularly helpful. In addition to enabling a better understanding of French dictionaries, DiCo may be useful to linguists interested in lexicology and diachronic and diatopic variation. Finally, it might also prove useful for building lexicons for natural language processing.

Keywords: metalexicography, French dictionaries, editorial policies, neologisms, Anglicisms

1. Introduction

This article presents DiCo, a database that records changes in the macrostructure of four French dictionaries and provides, for each modification, information on the microstructural level. This work originated in a study of spelling variation in dictionaries (Martinez 2009a), for which a manual comparison of the successive editions of printed dictionaries was necessary. This comparison was pursued beyond the scope of the initial study and became more systematic and exhaustive. Macro- and microstructural information from an extended list of dictionaries was recorded in a database. The list of new words and words deleted from the dictionaries was then published each year as individual web pages. The current paper presents the resource now made available in the form of a single downloadable document that includes additional information, as well as a browsable online version. Beyond describing the resource, the aim of this paper is to show how DiCo can be used to address a number of specific research questions.

The method of manual comparison of dictionaries used to build DiCo is described by Martinez (2009b, 2013). The two papers show that such a comparison can contribute to revealing how – and to what extent – dictionaries are updated. The French and English lexicographic landscapes have evolved in different ways.[1] This is also the case for the communication policies of publishing houses directed to the general public: while English publishing houses describe in detail the whole lexicographic process in dictionary prefaces, websites, blogs, etc., neither the front matter of French dictionaries nor the occasional press releases written by publishers say much about the real nature of the information included in the dictionaries and how the dictionaries are built. A metalexicographical investigation is therefore necessary to achieve a better understanding of such dictionaries and learn more about their content. This is where the DiCo database is particularly relevant. Metalexicographical studies, whether qualitative or quantitative, are usually based on the

manual analysis of a small sample of dictionary articles. DiCo enables quantitative observations of dictionary changes that occurred over an extended time span. These observations may, in turn, be the starting point for qualitative studies.

The content of the resource and encoding choices are described in Section 2. In Section 3, we exemplify how DiCo can be used in metalexicographical studies to describe the diachronic evolution of a given dictionary, or to confront lexicographic discourse with dictionary content. The studies presented in Section 3 show possible research approaches that rely solely on information found in the DiCo database. This section also illustrates how DiCo can be used to select relevant data to be further investigated, either by looking up dictionary definitions, etymologies and paratext or by comparing data to the content of other dictionaries.

2. Resource description

2.1 Dictionaries under study

The dictionaries included in the DiCo database are listed in Table 1. The *Dictionnaire de l'Académie française* is a multivolume dictionary written by an authoritative governmental institution; the first edition dates back to the late seventeenth century. DiCo provides information about the 8th edition (of which the A-G volume was published in 1932 and the H-Z volume in 1935) and the ongoing 9th edition (of which the first volume was published in 1992). The *Petit Larousse* (first published in 1905) and the *Petit Robert* (first published in 1967) are general-purpose single-volume dictionaries. Since 1997, they have both been published on a yearly basis – generally by late spring – and their front covers currently mention the year following the publication date, referred to as the *millésime* 'vintage'. In this paper, we use the word *edition* to refer both to major reworkings such as the different editions of the *Dictionnaire de l'Académie française* and to yearly updates (*millésimes*), such as those of the *Petit Larousse* and *Petit Robert*. The comparison of dictionaries initially started with the 2005 and 2006 editions of the *Petit Larousse* and was pursued, together with the comparison of the *Petit Robert* editions, backwards until 1997 and forwards until the present. The first editions of the *Petit Larousse* were then added to the study, starting from 1906. We now intend to bridge the gap between 1925 and 1997, depending on our ability to find copies of the missing editions. The *Dictionnaire Hachette*, first published in 1980, is a competitor to the latter two dictionaries. As the *Petit Robert* and *Petit Larousse* are the most comparable dictionaries under study and have the highest number of editions compared in DiCo, the observations in Section 3 focus on these two dictionaries.

Table 1: Dictionaries under study.

Abbreviation	**Name**	**Editions**
DAF	*Dictionnaire de l'Académie française*	8th ed. (1932-1935) and 9th ed. (ongoing)
DH	*Dictionnaire Hachette*	2017-2018
PL	*Petit Larousse*	1906-1925 and 1997-2020
PR	*Petit Robert*	1997-2020

2.2 Versions of DiCo

The DiCo database is intended primarily for metalexicographical and linguistic studies, but we believe that it will prove to be versatile and useful to a wide audience, including natural language processing (NLP) developers and language teachers. Relying on DiCo labels can help linguists conduct research in lexicology and terminology. For example, Sajous et al. (2020a) used DiCo to study the vocabulary of computer science. It can also be used in classes of French as a foreign language to select specific subsets of substandard vocabulary according to the targeted level of the learners, as suggested by Fievet and Podhorná-Polická (2011). Last, DiCo can be used to build lexicons relevant to NLP and corpus annotation. Attitude labels may help build sentiment lexicons; specialised terms may be used as seed words in a topic crawler such as BootCat (Baroni and Bernardini 2004) to build specialised corpora; and diatopic labels make it possible to build lexicons reflecting regional variations that are used for author profiling (Rangel et al. 2017) or to detect closely related languages (Tiedemann and Ljubešić 2012). These automatic approaches usually learn discriminative words from parallel or comparable corpora. The DiCo list of diatopic variants is ready for use and makes it possible to implement these methods even when no satisfactory corpus is available.

Because metalexicographers, linguists, language teachers and NLP developers do not share the same interests nor have the same background, we designed two versions of the database that differ in how the linguistic labels are reported, as detailed in Section 2.4.2. The two versions of DiCo are available for download[2] under a free licence as spreadsheet documents. The version intended for the general public is also browsable via an online user interface represented in Figure 1. This interface enables a user to sort the database and to filter the entries by the value of the fields described in Sections 2.3 and 2.4. The selected entries can be exported as a spreadsheet.

Colonne : Opérateur : est Valeur : valeur du filtre Effacer le filtre

Ch.	Type d'entrée	Dico	Année	No	Entrée	Form.	Équiv.	Catég.	Diatop.	Diatech.	Diachr.	Diafréq.	Attitude	Dianorm.	Wikt.
	filtre colonne...	PR	filtr		filtre colonne...		filtre colonne...	filtre colo	filtre color	filtre color	filtre colo	filtre colon	filtre colon	filtre colonn	filtre color
E	variante	PR	2020		GPA	sigle	gestation pour autr…	n.f.							2013-02-03
E	renvoi	PR	2020		Pâques → pâque										2012-12-18
E		PR	2020		affûté, ée /affut…		fin, vif	adj.							2007-02-08
E		PR	2020		agrobusiness		agrobizness	n.m.						anglicisme	2016-06-07
E		PR	2020		amitieux, ieuse		amiteux	adj.	Belgique, …		vieilli				2005-05-14
E	entrée cachée (sous androgenèse)	PR	2020		androgénétique			adj.							
E		PR	2020		anticasseur ou …			adj.							2013-11-03
E		PR	2020		arrière-nièce			n.f.							2007-06-14
E		PR	2020		assimilationnis…			n.m.					didactique		2019-01-10
E	entrée cachée (sous assimilation…	PR	2020		assimilationniste			adj. et n.							2019-01-10
E		PR	2020		autophagie			n.f.		physiologie					2009-01-27
E		PR	2020		azuki		haricot rouge du ja…	n.m. / ap…							2010-06-21
E		PR	2020		baclofène			n.m.		biochimie, …					2014-03-04
E		PR	2020		beignerie			n.f.	Canada						2013-04-15
E		PR	2020		binarité			n.f.					didactique		2011-02-26
E		PR	2020		biocapacité			n.f.		écologie					2019-08-13
E		PR	2020		biomolécule			n.f.		biologie					2009-02-21
E		PR	2020		biérologie			n.f.							2011-09-24
E	entrée cachée (sous biérologie)	PR	2020		biérologue			n.							2011-09-24
E		PR	2020		blockchain		chaîne de blocs (R…	n.f.		informatique				anglicisme	2016-02-27
E	renvoi	PR	2020	2	blocus → bloque										2007-02-09
E		PR	2020		bloque		blocus n, blocus n.m.	n.f.	Belgique				familier		2007-02-09
E	entrée cachée (sous boboïser)	PR	2020		boboïsation			n.f.							2010-05-12
E		PR	2020		boboïser			v.trans. / …							2008-03-06

Figure 1: The DiCo browser: an online interface for browsing the DiCo database.

2.3 Macrostructure

Each change identified is characterised by a headword,[3] the dictionary in which the change occurred and its year of publication, the main type of change (addition or deletion, signalled by E for *entrée* and S for *sortie*), and the type of entry involved:

- *regular entry:* refers to the addition or deletion of a standard whole article. A regular entry is signalled by an empty value in the 'type of entry' field.
- *variant* (*variante* in DiCo): when the definition of an article is worded only by means of synonym(s), the headword is considered a variant of the synonym(s). For example, *LED* (Fig. 2) is, according to its etymology, an acronym of the English expansion *Light Emitting Diode* and, according to its definition, *diode électroluminescente* is the official French equivalent.
- *run-on entry* (*entrée cachée* in DiCo): refers to the description of a word that is added to or removed from the main article of another word, instead of being presented as a headword. A run-on entry may be accompanied by a description (e.g. a definition or usage example) or simply mentioned, as in Figure 3: the run-on adjective and noun *agoraphobe* 'agoraphobic, agoraphobe' are derivatives of *agoraphobie* 'agoraphobia'. As the meaning of *agoraphobe* is transparent to the reader who understands the definition of the headword *agoraphobie*, no further explanation is required. A run-on entry that is promoted to a regular entry is considered a split entry and therefore an addition (cf. *split entry* below). For example, the verb *gentrifier* 'gentrify' entered PR2018 as a run-on entry under the noun *gentrification*. The following year, *gentrifier* became a headword, with its own article.
- *merged entry* (*fusion* in DiCo): when two articles of a given edition merge into a single article in the following edition, the 'lost headword' is considered a deletion from the dictionary nomenclature. For instance, the adjective *1. vétérinaire* 'veterinary' and the noun *2. vétérinaire* 'veterinarian', found as separate entries in PL2011, merged in PL2012.
- *split entry* (*scission* in DiCo): when splitting an article from a given edition results in two articles in the next edition, the new article is considered an addition. For example, the noun *éolienne* 'wind turbine' and the adjective *éolien, ienne* 'aeolian' appeared in the same article in PR2014. In PR2015, they split into two distinct articles.
- *cross-reference* (*renvoi* in DiCo): indicates that the information about a word is to be found under the article of another word. Cross-references often redirect the reader from a spelling variant to an equivalent form. For example, the cross-reference **FLAUGNARDE** ▸ **FLOGNARDE** entered PR2010 simultaneously with the article **FLOGNARDE** ou **FLAUGNARDE** (a kind of clafoutis, often made with apples).

LED ou LED [lɛd] **n.f.** – 1977 ◇ acronyme anglais, de *Light Emitting Diode* (1968) ■ ANGLIC. Diode électroluminescente (recomm. offic.). *Des led* ou *des leds. Les LED consomment peu et ne chauffent pas.* – **APPOS.** *Lampes, ampoules LED.* ■ **HOM.** Laide (laid).

Figure 2: English expansion and official French equivalent of **LED ou LED** (PR2010).

AGORAPHOBIE [agɔʀafɔbi] **n. f.** – 1865 ◇ du grec *agora* « place » et *-phobie* DIDACT. ■ Phobie des espaces libres et des lieux publics. *« franchir sans agoraphobie l'espace creusé d'abîmes qui va de l'antichambre au petit salon »* ***PROUST.*** – **adj.** et **n.** **AGORAPHOBE.**

Figure 3: Run-on entry **AGORAPHOBE** in the article **AGORAPHOBIE** (PR2010).

The number of change and entry types from all the years covered for each dictionary is given in Table 2. The large number of changes (e.g. article additions) in the DAF is due mostly to the massive addition of technical terms,[4] in addition to the half-century that elapsed between the 8th and 9th editions. The most striking difference is the large number of entries deleted from the PL compared to the PR and, to a lesser extent, the higher number of articles merged. These figures are discussed in Section 3.2.

Table 2: Changes recorded in the DiCo database.

Change	Type of entry	DAF	DH	PL	PR
Addition	cross-reference	336	3	198	182
	regular entry	9,222	20	5,229	2,791
	run-on entry	3	1	88	391
	split entry	660	0	312	82
	variant	66	0	512	137
Deletion	cross-reference	40	2	272	60
	merged entry	53	0	638	8
	regular entry	440	3	4,625	17
	run-on entry	0	0	26	25
	variant	8	0	495	4

2.4 Microstructure

For each change identified in the macrostructure of a dictionary, a set of information on the microstructural level is included in DiCo:

- the part of speech of the headword. When several parts of speech are present, they are listed in the order of their appearance in the article;
- the plural form, when irregular, e.g. *futal* (slang word for 'trousers') → *futals*. The plural is also given for multiword expressions (mostly compounds), e.g. *fer-à-cheval* 'horseshoe' → *fers-à-cheval* 'horseshoes' and borrowings, e.g. *slash* → *slashs ou slashes* (all examples are taken from PL1998);
- the variant type: when a headword is a variant (e.g. abbreviation, initialism, French back slang) of another word, the type of variant may appear in the etymology or in the definition. For example, *blème* entered PR2000, where it is described as an apheresis of *problème* 'problem';
- equivalents: when a word is defined by one of its synonyms (cf. Section 2.3), this synonym is mentioned in DiCo. For instance, the expansion *bicycle moto x* defines the headword *BMX*, added in PR2010. The French *télévérité* entered PL1998, where it

defines the borrowing *reality show*, which it is intended to replace. Conversely, the English *buzz* entered PR2010 (with the meaning of 'information that people are talking about'), where it is defined by the official French equivalent *bouche à oreille*.

Other microstructural information provided in DiCo is further described in the following two subsections.

2.4.1. Date of first known attestation and date of inclusion in *Wiktionnaire*

In the PL, only 28% of the new articles over the 1998-2020 period provide an etymology. In the PR, this section is theoretically obligatory. In addition to word formation and origin, a PR etymology gives a date of first attestation (the PL never does). Etymologies are not reproduced in DiCo for copyright reasons (just as, obviously, definitions, examples, citations, etc., are not). However, DiCo reports the dates of first known attestations, as they appear in the dictionary. The inclusion dates of words in the *Wiktionnaire* nomenclature (*Wiktionnaire* is the French language edition of *Wiktionary*), taken from the WIND resource[5] are the only external information added to DiCo. The rationale is that, when the first known attestation of a neologism is not provided by the dictionaries under study, and when no satisfactory diachronic corpus is available for automatic detection, the inclusion date of this neologism in *Wiktionnaire* may provide a hint as to its period of appearance. Another motivation for including this information is the opportunity to compare the lexicographical delay between 'professional' and 'amateur' dictionaries (cf. Section 3.3).

2.4.2. Linguistic labels

Linguistic labels can be clues in metalexicographical investigations, as we illustrate in Sections 3.1.2 and 3.4. As discussed in Section 2.2, DiCo labels may be used for several other purposes by different categories of users. Two versions of the resource have therefore been produced to meet the needs of these different tasks and users. In the version dedicated to lexicographers and linguists, the label values have been reported as they stand and have been assigned to the eleven types of the typology devised by Hausmann et al. (1989),[6] chosen because it is finely tuned and widely used (Corbin and Gasiglia 2011; Vrbinc and Vrbinc 2017). In the version intended for the general public, some label values have been changed and some label categories have been merged, as explained below.

Equivalent labels that have different forms in different dictionaries (not to mention within the same edition of a given dictionary) have been homogenised. For example, a non-expert user may not be interested in the fact that the same geographic area is indicated by the diatopic labels *Réunion* in one dictionary and *La Réunion* in another, and that the field label relating to statistics appears in both the singular (*statistique*) and plural form (*statistiques*). The differences between some diachronic labels, such as *vieilli/vieillit* and *vieux/vx*, have also been neutralised, even if these labels have different meanings. This choice was made in order to ease the lookup process in the DiCo browser and the design of specific sublexicons. Some users may indeed wish only to distinguish standard from non-standard vocabulary, and most NLP applications often require coarse-grained categories.

We also merged categories that partly overlap or that contain labels that are no longer used in current lexicographic practice.[7] First, some categories contain labels that have very few occurrences and that are used quite indifferently, even if they theoretically indicate different kinds of linguistic variation. For instance, there are only two occurrences of *oral* 'spoken', both collocated with *familier* 'colloquial' and three occurrences of *écrit* 'written'. Although there is no exact correlation between the communication channel and the degree of formality (Koch and Oesterreicher 2001), we decided to merge the diamedial and diaphasic categories. Second, labels may be polysemous at a given time or undergo semantic changes over time (remember that DiCo covers a period that spans from 1905 to 2020), and their

meanings may relate to different categories, leaving the user puzzled. For example, the label *vulgaire* 'vulgar' has followed the same trajectory in French as in English, starting from the same ambivalent use described by Wild (2008). Currently used to denote offensive and obscene terms, this label, meaning 'plebeian', was primarily used in the past to relate to 'the common people' (not necessarily conveying a negative connotation) and, meaning 'non-technical', has also been used to contrast with scientific terms. Depending on the entries and on the year of the edition, this label can be assigned to the diastratic, diaphasic, diatextual (when contrasting with 'scientific') or diaevaluative category. Along similar lines, the diaintegrative and dianormative categories have been merged, as explained in Section 3.4.

Another issue is the categorisation of the stigmatising label *populaire* 'popular' (a stylistic label relating to 'low language' or a diastratic label relating to the lower working class), which has been criticised (Podhorná-Polická 2011). As far as the lexicon is concerned, is the notion of sociolect still relevant in modern society, or do words marked as such rather relate to specific (informal) communicative situations, potentially involving any speaker, whatever their social class? The answer may vary across time and space. Wild reported a neutral use in the 18th and 19th centuries. In the 1970s, Rey-Debove (1971: 91-93) acknowledged the notion of sociolect (which she called *langue sociale*) on the grounds that, as France is a highly centralised country, dialects are losing importance, but social classes nevertheless exist in France, and the language of the working class (called *la langue populaire* by Rey-Debove) is different from the language of the wealthier class. Assigning a word to the class of people that use it is, however, a complex issue. Alain Rey, according to Corbin and Gasiglia (2011), in the preface of the *Grand Robert* (2nd ed.), criticised the use of *populaire* when it is intended to mean *familier* 'colloquial'. He proposed using it to label usages 'that educated people disapprove of'. Still with regard to France, Lodge (1989: 442-443), cited by Abecassis (2008), pointed out that the label *populaire* seems particularly inappropriate, '*[...] especially on account of the fact that social classes in France are not clearly definable. It could be said that both* fam. *and* pop. *are stylistic rather than social indicators on the low/high continuum.',* Abecassis added. Although criticised and abandoned by many dictionaries, such as the PL, *populaire* is still used in the PR and in the DAF. Our aim here is to report what is in dictionaries, not to discuss the relevance of a given label value. However, regarding categories, should *populaire*, when present in dictionaries,[8] be assigned to the diastratic or diaphasic category?

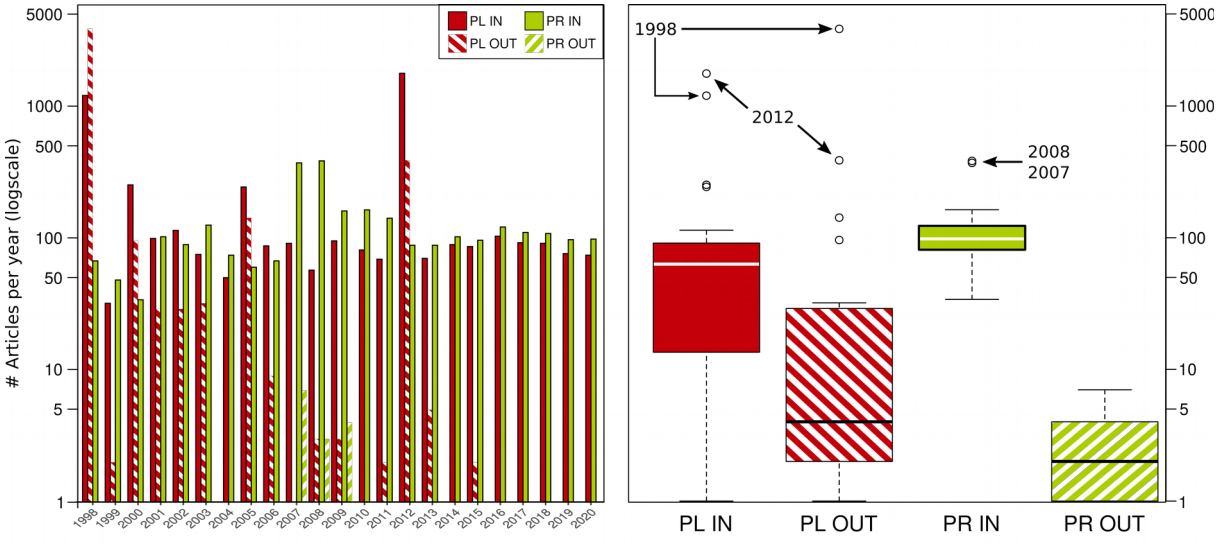
Assuming that a non-expert user with no strong theoretical and historical background in lexicography will easily grasp the notion of diachronic, diatopic or diatechnical labels but will not necessarily understand the discrete partition between other label categories, the diastratic, diamedial, diatextual and diaevaluative categories have been merged into a single broad category entitled *attitude*. According to Namatende-Sakwa (2011), who analysed the labelling practices in six monolingual English dictionaries, this broad category is close to the Macmillan Dictionary category entitled *Style and attitude labels*.[9] It also corresponds more or less to the category that Landau (2001) called *style, functional variety, or register* (specific labels corresponding to Landau's *taboo* and *insult* categories are almost never used in the dictionaries under study: there are only four occurrences of the label *injurieux* 'offensive', three of which are collocated with *vulgaire* 'vulgar' and one with the label *raciste* 'racist'). Our *attitude* category also encompasses the different *argot* 'slang' labels (*argot* 'slang', *argot militaire* 'military slang', *argot des prisons* 'prison slang', etc.), to which Landau (2001) dedicated a separate category all to itself. Again, depending on the words and dictionaries in question, such labels may be considered diastratic or diaphasic (or even field labels).

3. Description and comparison of dictionaries

We stated above that French publishing houses are very sparing in the information they communicate to the general public. Unlike the *Oxford English Dictionary*, whose complete list of new entries is published each time the dictionary is updated,[10] the Robert and Larousse publishers only mention a few buzzwords in occasional press releases. These releases comment only very briefly, if at all, on how new headwords are selected or how many articles are updated and do not mention the fact that information disappears from dictionaries (in particular when whole articles are deleted). A metalexicographical investigation is thus necessary to achieve a better understanding of such dictionaries and learn more about their content. Sections 3.1 to 3.3 illustrate how such investigations can be conducted on the sole basis of information included in DiCo.[11] The number of added, deleted and merged articles; the proportion of marked vocabulary and the most frequent labels; and the inclusion delays presented below were generated directly from the DiCo spreadsheet. In Section 3.4, we illustrate how DiCo can be used as a starting point and supplemented with additional material (e.g. data manually retrieved from the printed dictionaries) in order to conduct further investigations.

3.1 New articles

Figure 4 depicts the numbers of new regular articles added to, and articles deleted from, the PL and the PR over the 1998-2020 period. Figure 4a reports the raw numbers of additions and deletions for each year, while the boxplots depicted in Figure 4b represent their dispersion.[12]



(a) Raw Frequency **(b)** Variation

Figure 4: Number of additions and deletions of articles per year in the *Petit Robert* and *Petit Larousse* dictionaries.

A new edition of both dictionaries is published every year, generally characterised by a relatively stable and low number of article additions: with a median value of 100, the number of articles added yearly to the PR is slightly greater than that of the PL (median value of 69). Major reworkings are exceptions. Such reworkings occurred in 1998 and 2012 for the PL (1,198 and 1,764 regular articles added) and in 2007 for the PR (370 regular articles added and 383 the following year). Redesigns are also an opportunity for article deletions: 3,851 regular articles were removed from the PL in 1998 and 387 in 2012. The number of additions and deletions occurring during the 1998 and 2012 redesigns of the PL are identified as extreme values in Figure 4b. Deletions occurred regularly in the PL over the 1998-2015 period. Conversely, deletions occurred in the PR only in 2007 and the following two years,

with a total amounting to only 17 articles deleted (article deletion is further addressed in Section 3.2). Figure 4b shows that the PR and PL have comparable median values in terms of both article additions and deletions. However, the vertical stretch of the PL boxplots highlights its irregular rate of modification to the nomenclature, compared to the relatively stable rate of the PR. These observations raise a number of questions: What kind of vocabulary is added to French dictionaries every year? Are new articles all dedicated to neologisms? How are the words to be deleted selected? The following sections attempt to find out the answers.

3.1.1. Neologisms and lexicographical delay

For a word to make its way into a dictionary nomenclature, it must meet several criteria. Although the inclusion criteria depend on dictionaries' editorial policies, some are standard prerequisites, such as a sufficiently high corpus frequency. Another consensual criterion is the 'time endurance' of words: ephemeral words are not welcomed in dictionaries. Therefore, checking (manually or automatically) whether a word has established itself is possible only after a certain period of time has passed since its creation, which theoretically prevents the inclusion of recent neologisms in dictionaries. Dictionaries are even used as a corpus of exclusion for automated neology watch. However, the Robert and Larousse publishing houses boast the addition of recent buzzwords in their dictionaries.

As stated in Section 2.4.1, the PR provides the date of the first known attestation of a word. The variation in the time span between this date and the inclusion date of words in the PR is depicted in Figure 5 for each year from 1998 to 2020. The numerous extreme values (identified by circles) generally correspond to words attested for several centuries (the boxplot representation is not sensitive to such extreme values).[13]

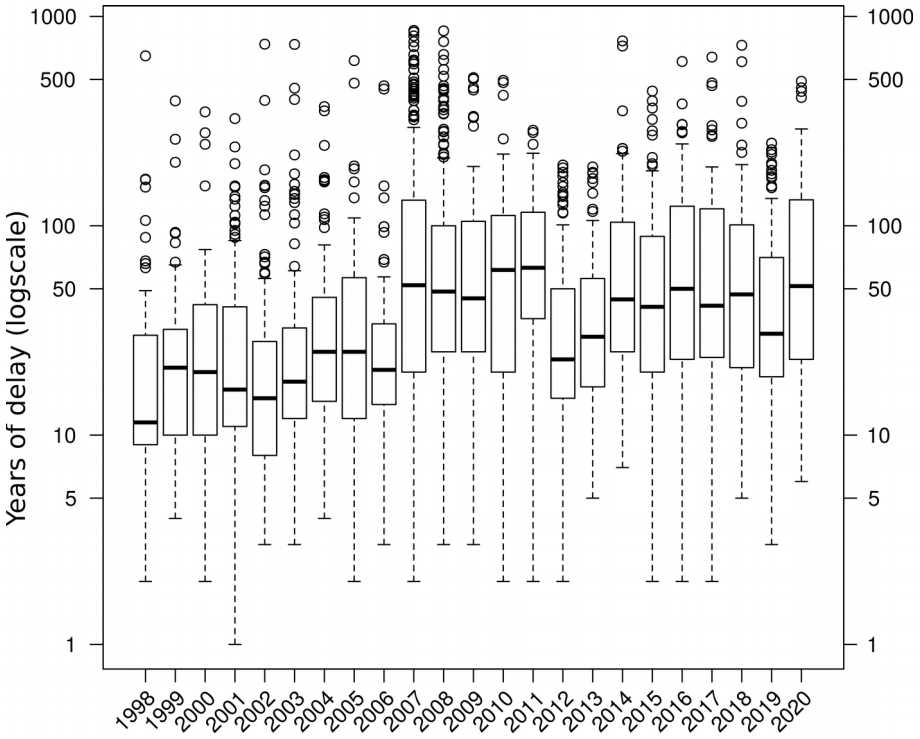


Figure 5: Delay between the first known attestation of words given by PR etymologies and their inclusion in the dictionary.

Before the 2007 reworking, the median delay ranged from 11.5 to 25 years (the average delay ranged from 36 to 57). Since 2007, the median delay has ranged from 23 to 63 years, with an average value of 46 to 109.5. Even if some words are included after only one year, new

articles are obviously not dedicated exclusively to neologisms. Section 3.1.2 investigates whether linguistic labels can reveal what kind of vocabulary is added every year to dictionaries and what kind of change(s) occurred in the PR in 2007.

3.1.2. Marked vocabulary

Diatechnical labels signal specialised terms that belong to a given domain. The proportion of specialised vocabulary among the new entries is given per dictionary for each year in Figure 6.

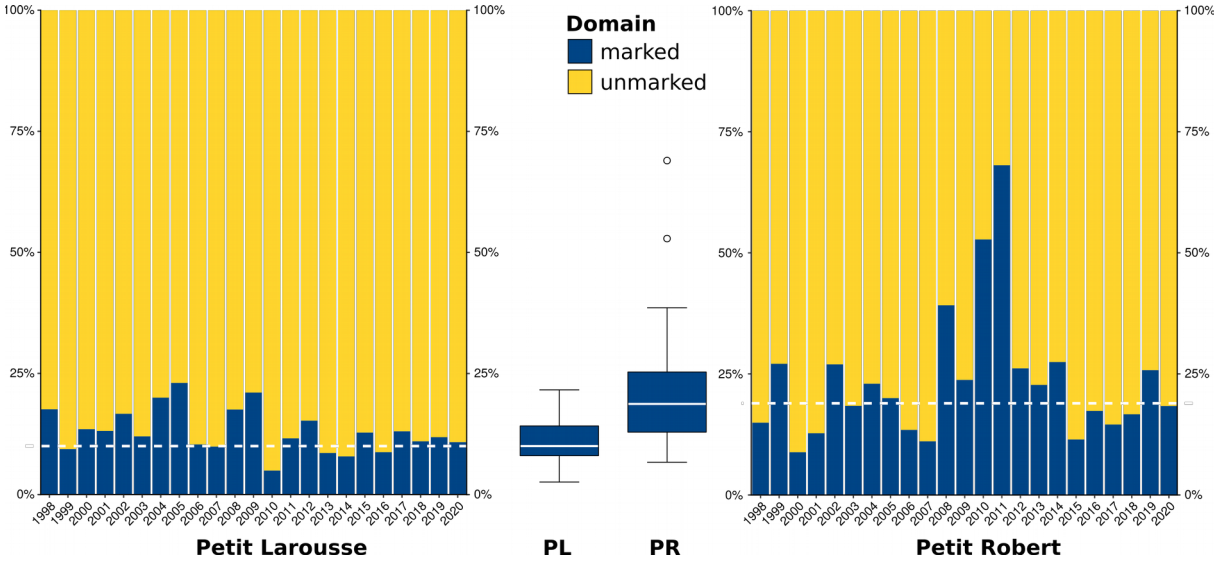


Figure 6: Proportion of specialised (marked) lexicon vs. general (unmarked) vocabulary in *Petit Larousse* and *Petit Robert* new entries.

This diagram shows that the PR includes more specialised terms than the PL (median values: 20% vs 12%). The stacked bar chart highlights that the rate is highly variable in the PR, ranging from 9% in 2000 to 68% in 2011. It also shows that the rate of domain-specific terms entering the PR increased after 2007. The most frequent domains in the PL and in the PR are given in Figure 7.

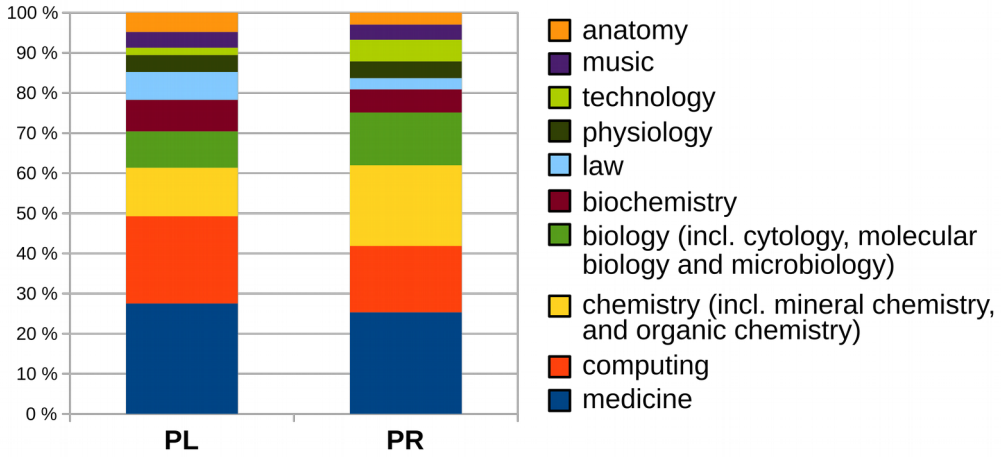


Figure 7: Ten most frequent domains in *Petit Robert* and *Petit Larousse* new entries.

In different proportions, the main domains are more or less the same in both dictionaries: medicine, computing, chemistry, biology, etc. For instance, the PL favours computing, while the PR includes more terms related to different subfields of chemistry. This statement should, however, be tempered. In a study on the computing domain, Sajous et al. (2020a) showed

inconsistencies in the PR diatechnical labelling: *podcaster* ‘to download a podcast’ is labelled, but *podcast* is not; *tchat* ‘chat (online discussion)’ is, but *tchatter* ‘to chat (online discussion)’ is not; etc.

Diatopic labels signal words used only in a given geographic area (state, region, etc.). As we did for domains, we report in Figure 8 the proportion of new vocabulary marked by a diatopic label per dictionary for each year.

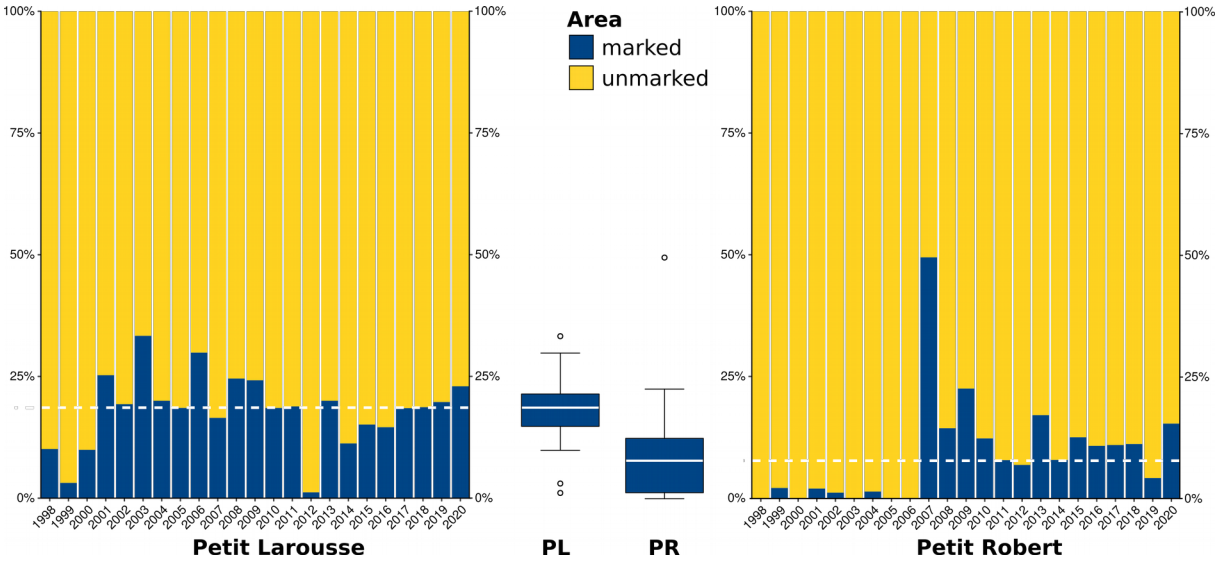


Figure 8: Proportion of new entries marked by a diatopic label.

We have seen that the PL includes fewer specialised terms than the PR. Conversely, the stacked bar charts and boxplots show that it includes more diatopic variants. Except for the 2012 and 1999 editions (year of a major reworking and year following another major reworking), which included only 1.2% and 3.1% diatopic variants, respectively, the PL rate has remained stable, hovering around a median value of 18.7% (compared to 7.8% for the PR). Thus, the PR favours specialised vocabulary rather than diatopic variants. Interestingly, the addition of diatopic variants has occurred mainly since 2007. The increase in the number of technical terms and diatopic variants added to the PR since the 2007 reworking explains the shift in lexicographical delay observed from this year to the present (described in Section 3.1.1): when the Robert publishing house opened its dictionary to words from the Francophonie, or suddenly and massively added terms from a given domain (e.g. chemistry), the contingent of new articles started to include words that had long existed, thus inducing a rise in the median age of new entries.

The most frequent geographic areas found in the PL and PR are given in Figure 9. They comprise regional France, overseas departments and territories or countries that are part of Francophonie. Regional France labels correspond to different granularities: towns (e.g. *Marseille* and *Sète* in the PR and *Lyon* in the PL), administrative subdivisions such as departments (e.g. *Gironde*) and regions (e.g. *Bretagne*), and historical and tourist regions (e.g. *Périgord* and *Anjou*). They are grouped under the ‘regional France’ category. Canada and Quebec have separate labels in both the PR and the PL, with the latter also distinguishing Acadia. The two dictionaries have the same general labels for Maghreb and Africa but use a different set of subdivisions that partly overlap: both have specific labels for Morocco and Algeria, but Black Africa is found only in the PR and Central and West Africa only in the PL. All specific African countries (other than those in the Maghreb) have fewer than four occurrences each. In the diagram, labels related to Africa are divided into the categories *Maghreb* and *Africa*.

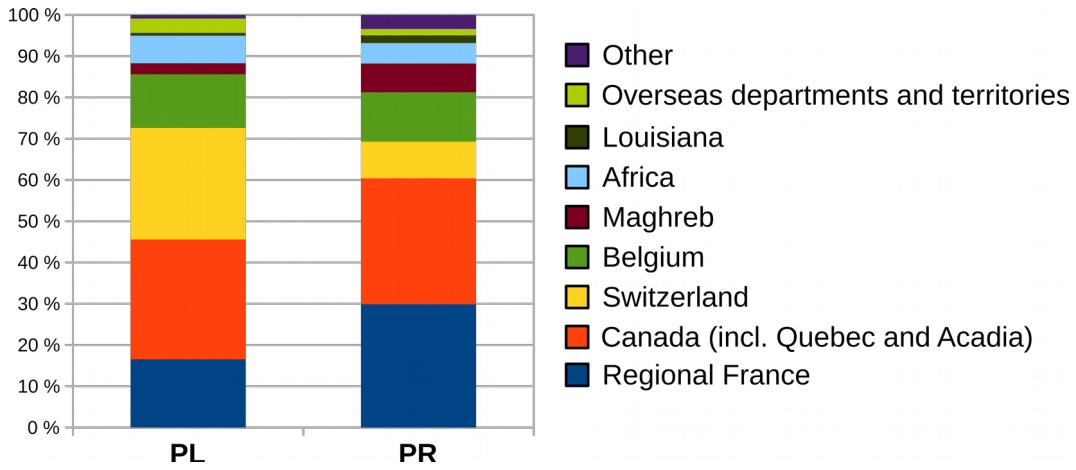


Figure 9: Distribution of diatopic labels in *Petit Robert* and *Petit Larousse* new entries.

Unsurprisingly, the areas ranking first are countries where French is one of the official languages: Belgium, Switzerland and Canada. Although the three main countries almost systematically rank first, lexicographic work may occasionally focus on a particular area. For example, 83% of the 'Francophonie words' entering PR2009 come from the Maghreb (*Maghreb*, *Algérie* 'Algeria' and *Maroc* 'Morocco' labels). The most striking difference in terms of proportion is the greater number of Swiss variants in the PL: while both dictionaries include, for example, the same proportion of Canadian words (29% of diatopic variants added to the PL, 30.5% to the PR), words from Switzerland represent 27% of the diatopic variants added to the PL as opposed to 8.9% in the PR (which favours regional variants from metropolitan France instead). Again, the explanation lies in redesigns: even if Swiss variants have been regularly added to the PL, 63% were added during its 1998 reworking.

3.2 Deletion and merging of articles

The publishing houses, in addition to providing highly selective information about new entries, never comment on the deletion of articles. Only 17 articles were removed from the PR from 1998 to 2020, whereas the PL deleted 4,594 regular articles during the same time span. These deletions normally relate to dated and rare words. However, only 8.7% of the entries deleted from the PL are marked by a diachronic label (e.g. *vieilli* 'dated'), and 2.7% are identified as *rare* words. The proportion of deletions of specialised terms was 10% higher than that of additions (28.3% vs 18%). Conversely, the rate of deleted words marked by a diatopic label was lower than that of new entries (2.5% vs 7.8%). The majority of deleted articles are not marked by any label: 2,634 of such unmarked entries represent 57.3% of article deletions. Caution should be exercised when interpreting these figures. One should not hastily conclude that 'regular vocabulary' is removed from the PL whereas dated words are added (cf. Section 3.1.1). Observations tend rather to suggest that the labelling of entries is far from systematic. For example, *arrosement* 'watering, irrigation' was not marked when it was removed from the PL in 2012. It is, however, a dated (marked as such in the PR) equivalent of *arrosage*. *Effaneuse*, a tool for stripping the top leaves from potato plants before harvesting, is marked by the domain label AGRIC. (agriculture), but *effanure* (the top leaf removed with that tool), deleted the same year as *effaneuse*, is not marked.

Regarding the merging of articles in the PL, 88% (561 out of 638) occurred during a major reworking (508 in 1998 and 53 in 2012). Few of them correspond to the lumping of variants, such as *budgéter* 'to budget', which merged with *budgétiser* in the PL1998 article **BUDGÉTISER** ou **BUDGÉTER**. The majority of article mergers are due to derivative words having different parts of speech but related meanings (e.g. the noun and adjective *vétérinaire* 'veterinary/veterinarian') and to words presented as homographs in one edition but treated as

a polysemous entry in the following edition. Such choices, whether linguistically motivated or due to the need to save space, are unstable: it is not unusual to observe lumping-splitting-lumping (or, conversely, splitting-lumping-splitting) cycles. For example, the polysemous noun *déferlante* (sense 1: ‘breaker’ and, by metaphor, sense 2: ‘overwhelmingly growing phenomenon’) merged in 1998 with the adjective *déferlant, e* ‘breaking’. The new article was split in 2005 into *déferlant, e* (adjective ‘breaking’ and noun ‘breaker’) and *déferlante* (noun ‘overwhelmingly growing phenomenon’). These two articles merged again in 2012, more or less returning to the 1998 configuration: a single article containing one section related to the adjective and another to the noun, but this time, the ‘breaker’ sense is duplicated and appears in the two sections. The section dedicated to the adjective, which reads *vague déferlante ou déferlante* ‘breaking wave or breaker’, also describes the noun: *vague qui déferle* ‘wave that breaks’.

3.3 Comparison with *Wiktionnaire* nomenclature

Previous studies have shown that lexical resources based on *Wiktionnaire* have better corpus coverage of French vocabulary than other existing resources (Hathout et al. 2014; Sajous et al. 2014). Sajous et al. (2018) showed that, in 2017, the new PR entries had entered *Wiktionnaire* 7.5 years earlier (median value). The DiCo database enables us to compare the year of inclusion of both PR and PL new entries to that of *Wiktionnaire* over an extended period. *Wiktionnaire* was launched in December 2003 and started being fed in late 2004. For each year since 2004, Figure 10 gives the proportion of PL and PR new entries that were already included – at that time – in *Wiktionnaire*. The inclusion dates of words in the *Wiktionnaire* nomenclature were taken from the WIND resource (cf. Section 2.4.1). A new word entering the PL/PR is counted as present if it entered *Wiktionnaire* before its inclusion date in the PL/PR.[14] Conversely, it is counted as absent if it is not (currently) in the *Wiktionnaire* nomenclature or if it entered *Wiktionnaire* after it entered the PL/PR. Of course, as *Wiktionnaire* was empty in 2004, the rise of the rate of presence of words in this dictionary (observed in the left-hand part of the diagram) informs more about its filling process at that time than about the professional dictionaries. The right-hand part of the curves is more telling both about the differences between the PL and the PR and the irregularities occurring for each of them. First, the curves show that the PL has more ‘specific’ new entries (that were not yet recorded in *Wiktionnaire*) than the PR. Second, two observations deviate from the general trend of the curve, showing that something unusual happened at some point. Although the rate of presence in *Wiktionnaire* is on the rise (38.9% in 2010 to 65.5% in 2012), the PR value of 31.4% in 2011 seems particularly low. If we examine the 37 articles added to the PR that were absent from *Wiktionnaire*, 28 of them (76%) are labelled as belonging to related scientific domains: *biologie* ‘biology’, *biologie moléculaire* ‘molecular biology’, *biologie cellulaire* ‘cytology’, *biochimie* ‘biochemistry’, etc. An explanation can be found in a review of the 2011 edition written by Martinez (2010): terms in the field of biology and related domains represented more than 40% of article additions that year. Other additions that are missing from *Wiktionnaire* include 3 Belgian variants, 2 prefixes and 4 unmarked regular words. The 2019 downshift of the PL (70.7% of presence in *Wiktionnaire* as opposed to 81.1% in 2018 and 86.5% in 2020) is less dramatic. The 13 words missing from *Wiktionnaire* are 4 diatopic variants (from Morocco, Belgium, Alsace and Lyon), one borrowing (*open access*) and one pseudo-Anglicism (*mapping vidéo*) that both have French equivalents, one borrowing from Japanese (*teppanyaki*) and 6 other neologisms.

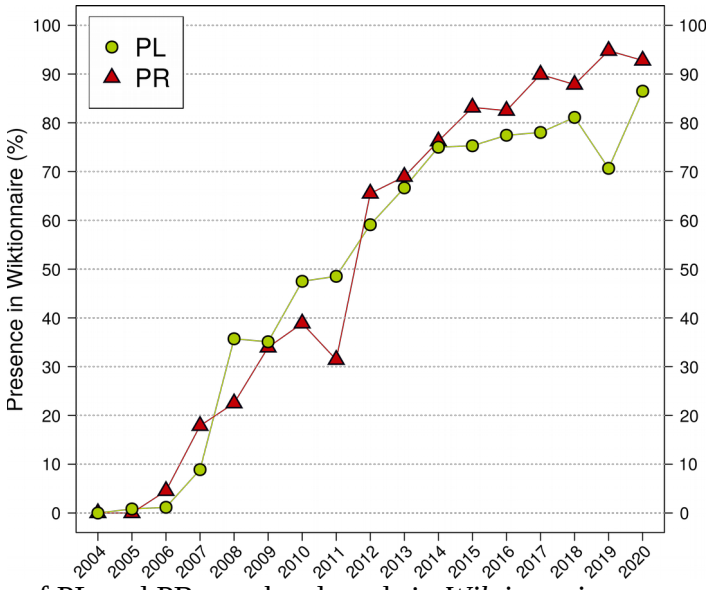


Figure 10: Presence of PL and PR new headwords in *Wiktionnaire* nomenclature.

For the words added to the PL and PR that are also present in *Wiktionnaire*, Figure 11 shows, for each year, the variation in the delay between the date of inclusion of the new words in the two professional dictionaries and in *Wiktionnaire*. Negative delays correspond to words that were first included in the PL or in the PR and later entered *Wiktionnaire*; positive delays correspond to the elapsed time between the inclusion of the words in *Wiktionnaire* and their subsequent inclusion in the commercial dictionaries.

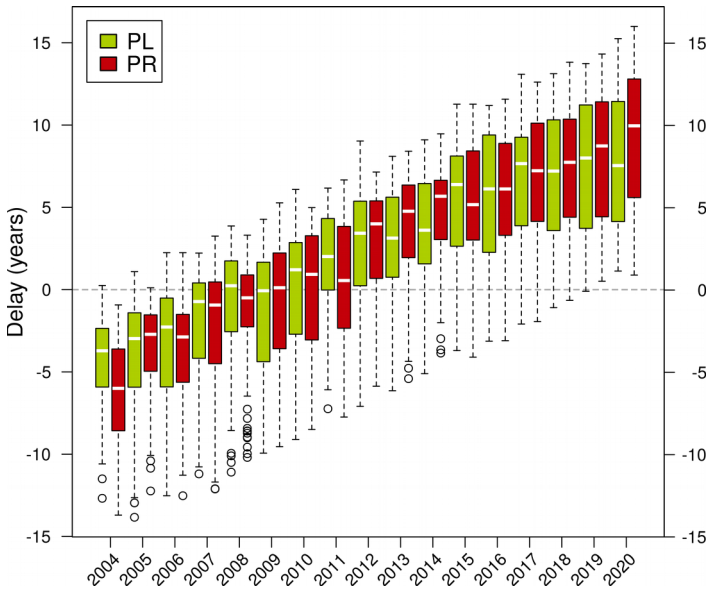


Figure 11: Inclusion delay of PL and PR new headwords compared to that of *Wiktionnaire* nomenclature.

Not only does the percentage of PL and PR new entries already present in *Wiktionnaire* increase (cf. Fig. 10), but the gap between the date of inclusion in *Wiktionnaire* and in the professional dictionaries also widens (this trend is slightly stronger in the PR for the last three editions). It seems reasonable to believe that *Wiktionnaire* has caught up with the 'core lexicon' and that new records are mostly current neologisms. To conclude on PL and PR, the increasing inclusion delay observed for the commercial dictionaries with respect to

Wiktionnaire nomenclature indicates that the new additions to PL and PR are more related to remedial treatments involving diatopic variants or standard well-established words than to the inclusion of true recent neologisms.

3.4 **Origin of headwords: diaintegrative or dianormative label?**

As mentioned in Section 2.4.2, we decided to merge the diaintegrative and dianormative label categories. In this section, we explain the rationale for this choice. Let us start with the following observation: the selection of new words for PR2020 included, in the culinary domain, the Japanese *azuki*, *ramen* and *soba* without labelling these words, but the English *welsh* 'Welsh rarebit/rabbit' entered the same edition with the ANGLIC. label. This raises the following questions: Does the diaintegrative label found in the PR and the PL only provide information about the origin of words? What criteria are used by French dictionaries to assign a diaintegrative label to borrowings? Are these criteria linguistically motivated? The current section intends to answer these questions in light of linguistic considerations and lexicographic discourses found in dictionary paratext. To establish a list of criteria that could explain the labelling of words as Anglicisms, we delineate the notion of Anglicism by considering two competing definitions found in the literature and examining how the PR and PL describe the corresponding label. We also present the French language policy that can influence dictionary marking practices. Then, we describe how we selected a sample of data from DiCo to assess the ability of the inventoried criteria to predict label attribution. Last, we apply the identified criteria to the selected entries and discuss the presence or absence of a label.

3.4.1. Definitions of *Anglicism*

Saugera (2017: 42-43) pointed out that there is no consensus in the literature on the criteria for classifying a word as an Anglicism and that two working definitions compete: one based on etymology, which classifies as an Anglicism any form that can be historically documented as stemming from the English language, and another based on native speakers' recognition of English words in the recipient language (spelling, morphology, etc.). Both definitions could also apply to any other donor language. The definition based on formal appearance, relying on individual perception and knowledge of the donor language, is necessarily subjective. Moreover, it fails to identify borrowings from other foreign languages via English as a medium of transmission (e.g. French *pastrami* borrowed from English *pastrami*, itself stemming from Yiddish *pastrame,* is not readily recognised as an Anglicism). The etymology-based definition could easily be implemented in the PR, as the dictionary systematically mentions the donor language of the etymon. However, ANGLIC. is the only diaintegrative label in the PR. It is found in addition to the origin of the word given in its etymology: *mot anglais* 'English word'. Like Anglicisms, borrowings from other origins are signalled in their etymologies (*mot espagnol* 'Spanish word', *mot allemand* 'German word', etc.), but, conversely, their definitions do not contain any diaintegrative label (e.g. *Hispanism* or *Germanism*).[15] The special treatment of Anglicisms suggests that English has a special status among donor languages. However, not all English borrowings are labelled Anglicisms, as illustrated below. The etymology-based definition is therefore not sufficient to explain the ANGLIC. label. We discuss the definition based on formal appearance in Section 3.4.7.

3.4.2. Paratext

If we have a look at the PR list of abbreviations, ANGLIC. is defined as follows: '*mot anglais, de quelque provenance qu'il soit, employé en français et critiqué comme emprunt abusif ou inutile (les mots anglais employés depuis longtemps et normalement en français ne sont pas précédés de cette marque)'*. Thus, an Anglicism is, in the PR, an English word, whatever its origin, used in French and criticised as an improper or unnecessary borrowing. Conversely,

English words that have long been used ‘normally’ are said to be unlabelled. This definition alone might suffice to demonstrate the prescriptive nature of the ANGLIC. label in the PR: in this dictionary, ANGLIC. does not mean ‘borrowing of English origin’ but rather means ‘criticised Anglicism’. The exact meaning of ‘normally used’ and ‘improper or unnecessary’ is not further commented on, but an explanation is found in the preface, which states that the number of Anglicisms is higher than that of other foreign words, even if a substantial inflow of words borrowed from Italian, Arabic, Spanish, German, Japanese and Russian is apparent. Borrowings are justified, the PR explains, by the need to name ‘things coming from afar and that had remained ignored’. According to the dictionary, some Anglicisms are more objectionable than others because they are unnecessary: the prestige of the United States, its economic power and its technoscientific leadership generate a flood of borrowings even when appropriate French equivalents exist. By way of comparison, the PL list of abbreviations stipulates that the ANGLIC. label means Anglicism (with no mention of criticised Anglicisms) and the short preface reads: ‘*Nous favorisons l’usage, lorsqu’il est avéré [...] Pour les anglicismes, nous signalons les équivalents proposés par les autorités linguistiques’* (We favour established usage [...] For Anglicisms, we indicate the equivalents recommended by the linguistic authorities). The PL refers to equivalents established in usage and to those provided by linguistic authorities. Both can influence the marking of borrowings in dictionaries. We discuss below how official substitutes are coined by authoritative institutions. The criterion of the existence of equivalents belonging to the established usage is assessed in Section 3.4.5.

3.4.3. French language policy and official equivalents[16]

In France, authoritative institutions are in charge of implementing the French language policy. For instance, the task of the *Commission d’enrichissement de la langue française* is to coin and promote French equivalents for scientific and technical terms. The newly created words then have to be validated by the *Académie française*. Words that have received the go-ahead are published in the *Journal Officiel* ‘Official Journal’ and may be mentioned in dictionaries. In the PR articles dedicated to Anglicisms, these equivalents are signalled by the abbreviation ‘Recomm. offic.’ (cf. Fig. 2: *diode électroluminescente* is mentioned as the official equivalent of LED). Such official substitutes are a strong argument in favour of marking Anglicisms as criticised. For example, the two words (out of six) labelled Anglicisms that entered the PL over the 1998-2000 period with the additional *déconseillé* ‘discouraged’ qualifier are *firewall* (PL2005) and *digitalisation* (PL2012), which have official substitutes (*pare-feu* and *numérisation*). Equivalents are also coined for non-technical words, even if they are classified as domain-specific (sport, leisure, etc.): *à coûts réduits*, a substitute for *low-cost*, is classified in *économie et gestion d’entreprise* ‘economy and business management’; *atelier collaboratif* replaces *fablab* in *recherche-industrie* ‘research-industry’; *infox* replaces *fake news* in *communication*; etc.[17] Some others, such as *prêt-à-monter*, coined to replace *kit*, fall in the category *tous domaines* ‘all domains’. In the musical domain (ART/Musique category), *disc jockey* and *DJ* have two official substitutes: *platiniste* (after *platine*, another word for *tourne-disque* ‘record player’), officially adopted in 2011, refers to an artist who mixes different sources to produce an original creation, and *animateur*, officially adopted in 2020, refers to a person who plays one recording after another during a party. The PR mentions the official *animateur* in the entry *disque-jockey* (dated 1968, or 1954 with the spelling *disc-jockey*) and *platiniste* in the entry *DJ*. Although *DJ* has both the *animateur* and *platiniste* meanings according to the PR, *animateur* is not mentioned in this entry. *Animateur* is polysemous and already had an article that had been updated to describe the *disque-jockey* meaning. *Platiniste* is monosemous and very rare: 74 occurrences in frTenTen2017 (0.01 per million) and a few hits from Google corresponding to dictionary definitions. The PR denied this official coinage its own article. Still in the musical domain, the Italian *adagio*, *arioso*, *mezzo-soprano* or *a*

capella are not labelled. The lack of official equivalents for these words (the *Commission d'enrichissement* never felt the need to replace them) could explain the absence of a label. However, the dynamic markings *mezza voce*, *piano*, *forte* and their degrees *pianissimo* and *fortissimo* all have synonymic definitions in the PR: *à mi-voix*, *doucement*, *fort*, *très doucement*, *très fort* (*mezzo-piano*, *mezzo-forte*, *pianississimo* and *fortississimo* are absent from the nomenclature). Due to these French equivalents, although not officially stamped, the Italian adverbs could be labelled 'criticised Italianisms'. No label marks them, however.

3.4.4. Selection of a sample of borrowings from DiCo

The PL contains only 6 new words labelled Anglicisms over the 1998-2020 period, as opposed to 286 in the PR. Given the common use of the label in the PR and its scarcity in the PL, the remainder of Section 3.4 focuses on the PR.

English is known to provide French with numerous specialised terms, stemming for example from the domains of computing and the Internet. Examples from this area are examined in Section 3.4.8. To compare borrowings originating from different languages, we turned to another field, where English is not the exclusive provider. In the culinary domain, which supplies dictionaries with a significant proportion of new borrowings of various origins, English is an outsider. Examples of borrowings related to this domain that have been added to the PR are given in Table 3. The headwords were chosen by manually browsing DiCo so as to select different donor languages and years of inclusion. The information related to *cheeseburger*, *hamburger* and *hot-dog* was not recorded in DiCo (the words were included before 1998) and was manually retrieved from the printed dictionary. For all other headwords, the year of inclusion, first attestation and linguistic labels were taken from DiCo. The etymology section is not included in DiCo for copyright reasons (only the first attestation is stored), so the donor language was retrieved manually from the dictionary. The PL edition that included the borrowing is also reported (grey cells indicate the absence of the borrowing in this dictionary), as are the corresponding labels.

In the PR, grammatical labels indicate plural usage (AU PLUR.) for *ramen*, *soba* and *nacho*, and a field domain indicates the culinary domain (CUIS.) for *burrito* (which, interestingly, is the only word in the list marked as such). Apart from these labels, only words whose donor languages are English and American English are marked by a diaintegrative label. No borrowing from the list is labelled Anglicism in the PL.

In the following, we check in the PR whether the criteria for labelling words as Anglicisms, as identified above, consistently hold when confronted with evidence found in the dictionary: is the existence of a French equivalent – official or not – a good predictor of the presence of the ANGLIC. label? Conversely, is the entrenchment of an Anglicism in the French language (age, frequency and/or lexicalisation of the word: spelling adaptation, existence of derivatives, inflections, etc.) a predictor of the absence of this label? If not, do other features, unmentioned by the dictionary (e.g. the formal appearance of the borrowing or the nature of its referent), play a role in the labelling?

Table 3: Examples of inclusion of borrowings related to the culinary domain (sorted by donor language and PR edition year).

Headword	Donor language (PR)	Edition		First attestation (PR)	Labels	
		PR	PL		PR	PL
kémia	Algerian Arab	2017		1907	-	
cheeseburger	English	before 1998	1998	1972	ANGLIC.	-
hamburger	American English	before 1998	before 1998	1930	ANGLIC.	-
hot-dog	American English	before 1998	before 1998	1929	ANGLIC.	-
pastrami	American English < Yiddish < Romanian	2008	2012	1976	-	-
wrap	American English	2012	2017	1998	ANGLIC.	-
burger	American English	2015	before 1998	1982	ANGLIC.	-
barista	English < Italian	2015	2018	1993 in Canada	-	-
latte	English < Italian	2020		1998	-	
welsh	English	2020		1926	ANGLIC.	
waterzoï	Flemish	2001	before 1998	1765/1846	-	-
mezze	Greek, Turkish	2006	2005	1937	-	-
enchilada	Latin American Spanish[18]	2008		1990	-	
fajita	Latin American Spanish	2008		1994	-	
burrito	Latin American Spanish	2010		1987	CUIS.	
antipasti	Italian	1998	2015	1980	-	-
bruschetta	Italian	2015	2015	1991	-	-
focaccia	Italian	2016	2018	1832	-	-
ciabatta	Italian	2018	2017	1997	-	-
spritz	Italian < German	2018		1882; widespread in early 21st century	-	
azuki	Japanese	2020		1878; widespread around 2010	-	
ramen	Japanese	2020	2020	1985	AU PLUR.	-
soba	Japanese	2020		1954	AU PLUR.	
taco	Nahuatl	1998	before 1998	1988	-	CUIS.
tapas	Spanish	2010	before 1998	1987	-	CUIS.
lomo	Spanish	2016		1993	-	
piquillo	Spanish	2017	2015	1991	-	-
nacho	Spanish	2018		1990	AU PLUR.	
pad thaï	Thai	2017		1990	-	
massala	Urdu	2017		1902	-	

3.4.5. Age of words and existence of French equivalents

The Italian *spritz* and Japanese *azuki* have spread only since the early 21st century and around 2010, respectively, but both entered PR2020 without a label, whereas *hot-dog* and *hamburger* (dated 1929 and 1930) are (still) labelled. No one can seriously argue that the latter two words have not long been used adequately or that either of them has a French equivalent. We can speculate that *hot-dog* and *hamburger*, labelled when they entered the PR, should have lost their labels but that they still retain them because they have not been updated. Even though *wrap* and *welsh* are dated 1998 and 1926, respectively, they may not be widespread (except in Northern France for *welsh*, where it is a popular dish), which could justify their labels, according to the PR criteria. However, *burger* (apheresis of *hamburger*) entered the PR belatedly (in 2015) after having long been used (it is dated 1982) and having spread widely, both as a designatum (burgers are sold on every street corner, even in vegetarian versions) and as a denotatum (it has 41 times more occurrences than *azuki* in the frTenTen 2017 corpus), but is nevertheless labelled ANGLIC. On what basis can it be considered *abusif ou inutile* 'improper or unnecessary', given that no French equivalent exists? Conversely, *antipasti* is an assortment of hors d'oeuvres whose singular *antipasto*, mentioned in the *antipasti* etymology (but absent from the PR nomenclature), is said to be the translation of *hors d'œuvre*. The existence of a French equivalent apparently does not apply to consider *antipasti* an 'improper or unnecessary' borrowing. Just as in the musical domain (cf. Section 3.4.3), Italianisms that have a French equivalent are unlabelled, while English borrowings with no substitute are criticised.

3.4.6. Nature of the referent

According to the PR, *taco* (PR1998, dated 1988) is of Nahuatl origin (cf. Fig. 12). Like the Latin American Spanish e*nchilada* and *fajita*, it is not labelled. The three words are defined by meronymy and hypernymy relations involving *tortilla*. The English *wrap* entered the PR in 2012. In contrast to other tortilla-based meals, *wrap* is signalled by a diaintegrative label (cf. Fig. 13). It is exemplified by 'chicken and crudités wraps' (*taco* is exemplified by 'chicken tacos') and defined as a 'sandwich made of a wheat tortilla, rolled in the form of a cone, and stuffed'. Reading this example and this definition, one cannot help wondering what the difference is between a *wrap* and a *taco* apart from the shape (the taco tortilla is folded in two, according to the PR), the type of flour used (wheat or corn)… and the presence of the diaintegrative label. Thus, the nature of the referent seems to have nothing to do with the marking (nor does the existence of a French equivalent: neither *wrap* nor *taco* has a French equivalent). The American English origin of the headword is the only clue.

TACO [tako] **n. m.** – 1988 ◇ du nahuatl ■ Plat mexicain fait d'une galette de maïs **(➤ tortilla)**, pliée en deux et fourrée. *Des tacos* [takɔs] *au poulet.* ■ **HOM.** Tacaud, tacot.

Figure 12: Definition of *taco* (PR1998).

WRAP [vrap] **n. m.** – 1998 ◇ mot anglais-américain, de *to wrap* « envelopper » ■ ANGLIC. Sandwich composé d'une tortilla de blé roulée en forme de cornet et garnie. *Des wraps au poulet et crudités.*

Figure 13: Definition of *wrap* (PR2017).

3.4.7. Formal appearance

Let us consider the three unlabelled Anglicisms in Table 3: *pastrami, barista* and *latte*. *Pastrami* is a borrowing from American English but stems from Romanian via Yiddish. The PR etymology identifies *barista* as *mot anglais, de l'italien* 'English word, from Italian' (French *barista* < English < Italian). *Latte* follows the same borrowing pattern (French *latte* < English < Italian). The donor language is English, but the PR considers it an Italian word, as the – unusually worded – etymology suggests: *mot italien, par l'anglais* 'Italian word, via English'. French native speakers will probably not detect the Yiddish or Romanian origin of *pastrami* and possibly the Italian origin of *latte* and *barista*. They will probably not recognise these words as English either. Thus, the formal appearance (the perception of the donor language) may be an unmentioned criterion for labelling: borrowings recognised as English are criticised in the PR, whereas others are not. In the PR description of the ANGLIC. label, *de quelque provenance qu'il soit* 'whatever its origin', which qualifies *mot anglais* 'English word', should probably be nuanced.

3.4.8. Derivation of borrowings: a sign of lexicalisation?

Changes to the spelling, inflected forms or derivatives of borrowings could be signs of the entrenchment of these words in the recipient language. Apart from the different kinds of burgers, the borrowings taken from the culinary domain do not give rise to derivatives. We turned to examples taken from the computing and the Internet areas. Table 4 reports examples of 1) words that have been borrowed from English to French and the subsequent French derivatives and 2) bases and derivatives that have both been borrowed from English to French, according to the PR. The year of edition, first attestation, and labels as well as the official equivalents were taken from DiCo. The headwords in square brackets are alternative spellings that were added after word inclusion. The bases of the derivatives and the etymons of the borrowings (and in the latter case the donor language) were retrieved from the printed dictionary. An indication of the headword meanings (often transparent) is given by the authors, based on the PR definition.

The noun *chat* '(online) chat' entered PR2002. The verb *chatter* 'to chat (online)' and the noun *(t)chatteur* 'person involved in a chat' entered the next edition. *Chatter* is a denominal verb of the French *chat*, according to the PR. *Chat* is criticised, while *chatter* is not, possibly due to the existence of an official substitute for the former and not for the latter. The *-eur* suffix and the change in spelling, reflecting the pronunciation of *tchatteur* (without an initial *t*, *chatteur* is pronounced /ʃatœʀ/), do not explain the absence of a label for this word: If they did, the nouns *dealer* and *hacker*, for which masculine and feminine inflections also exist under the forms *dealeur, euse* and *hackeur, euse* would not be labelled either. Neither *troll* nor *troller* 'to troll' has a French equivalent, but *troll* is criticised, while *troller* is not. No conclusion can be drawn, however, regarding the influence of morphological derivation: *troller* is said to be coined from the English *to troll* rather than being a derivative of the French *troll*. *Hackeur* has no equivalent and derives from the French *hacker*, for which the official substitute is *fouineur* (the unofficial *pirate* has established itself instead). Both are criticised. *Dealeur, dealer* and *deal*, dated 1970-1980, are all criticised, although only *deal* has a French equivalent. Surprisingly, none is a derivative from an existing French Anglicism: all have been directly borrowed from English, according to the PR. Conversely, *tweeter* and *retweeter* derive from the French *tweet*, borrowed from English. All of them are criticised despite the lack of equivalents. *Podcaster* entered the PR before the French *podcast*, from which it derives. *Podcaster* is not criticised, but *podcast* is (the official equivalent corresponds to only one sense relating to the process – broadcasting –, rather than the product, i.e. the audio/video file). No clear pattern emerges from these examples. In the PR, the production of derivatives is not considered a sign of the entrenchment of a borrowing, serving as a base in

the recipient language, which could justify the deletion of the normative label. For the three uncriticised words found in Table 4, either they are said to be direct borrowings (*troller* does not derive from the French *troll*) or their base has an equivalent, while the derivative has not (*chat*/*chatter* and *chat*/*chatteur, euse*). Instead of shedding light on the criteria for labelling a word ANGLIC., these examples raise new questions: On what basis does the PR decide whether a word derives from an Anglicism or whether it is borrowed directly from English? How consistent is the labelling practice in the PR? The latter question involves more than the marking of Anglicisms: regarding the field labels, *chat* is labelled INFORM. (*informatique* 'computing'), but *chatter* is not; *troll* is, but *troller* is not (and *burrito* is the only word labelled CUIS. out of the 30 shown in Table 3).

Table 4: PR borrowings and morphological families.

Headword	**Meaning**	**Edition**	**First attest.**	**Labels**	**Base or etymon**		**Official equivalent**
chat	(online) chat	2002	1997	ANGLIC. INFORM.	EN	chat	causette (now: dialogue en ligne)[19]
chatter	to chat (online)	2003	1998	-	FR	chat	-
chatteur, euse [ou tchatteur, euse]	person involved in a chat	2003	1998	-	FR	chat	-
[dealeur, euse ou] dealer	drug dealer	before 1998	1970	ANGLIC.	EN	drug dealer	-
dealer	to deal	before 1998	1980	ANGLIC. FAM.	EN	to deal	-
deal	deal	2008	1980	ANGLIC. FAM.	EN	deal	accord, négociation, contrat
geek	geek	2010	1996	ANGLIC. FAM.	EN	geek	-
geeker	to act like a geek	2017	2001	ANGLIC. FAM.	FR	geek	-
[hackeur, euse ou] hacker	hacker	2002	1984	ANGLIC.	EN	hacker	fouineur
hacker	to hack	2017	1995	ANGLIC.	FR	hackeur	-
podcast	podcast	2009	2004	ANGLIC.	Am EN	(i)Pod + (broad) cast	diffusion pour baladeur[20]
podcaster	to download a podcast	2008	2005	INFORM.	FR	podcast	-
troll	troll	2015	2005	ANGLIC. INFORM.	EN	troll or to troll	-
troller	to troll	2017	2008	-	EN	to troll	-
tweet	tweet	2012	2009	ANGLIC.	EN	tweet	-
tweeter	to tweet	2012	2009	ANGLIC.	FR	tweet	-
retweeter	to retweet	2018	2009	ANGLIC.	FR	re + tweeter	-

3.4.9. Discussion

We have shown that the availability or lack of a French equivalent (whether official or not) is not a satisfactory explanation for the presence or absence of a diaintegrative label in the PR. Neither is the novelty or the entrenchment of a borrowing (age, frequency, production of derivatives and inflections) or the nature of its referent. The examination of examples reveals more inconsistencies in the marking than it explains the ANGLIC. label: *hello* (dated 1895) is a criticised Anglicism, while *bye-bye* (dated 1934) is colloquial but not criticised. In the musical domain, both *ska* and *groover* 'to groove' entered PR2015. They are criticised and not criticised, respectively. Thus, not only do Anglicisms suffer unfair treatment compared to words from other origins, but injustice also occurs between English words. Apart from labelling inconsistencies, the only reliable predictor of the presence of a diaintegrative label is the origin of the headword: English and American English words deserve the label (which can now be called dianormative) unless other primary donor languages hide their English origin. Non-English words do not, even when they are new, are rare, and have direct translations. The dictionary data demonstrate a strong discrepancy between the definition of the ANGLIC. label, the discourse in the preface and lexicographic practice.

The attitude of French lexicographers towards Anglicisms may have paralleled, to some extent, shifts in public opinion. While English borrowings were a symbol of modernity, seen positively by part of the population in the 1960s and 1970s, the number of opponents of 'Anglomania' grew.[21] According to Saugera (2017: 3-5), the image of Anglicisms as lexical polluters was shaped partly by purists and institutions, such as the *Académie française*, which received preeminent media attention. The *Académie*, which Saugera called the 'French words police', guards against English loanwords. Its approach to Anglicisms 'does not allow a linguistic case to be appreciated objectively; it treats borrowings not as linguistic data but as targets for eradication' (2017: 141). The *Académie* writes its own dictionary, which reflects this attitude. To what extent should the *Académie* vision influence other lexicographers? Landau (2001: 231) wrote (about taboo words) that 'the moment the lexicographer accedes to the principle of excluding any words on the grounds of someone else's taste, he has relinquished control of his dictionary and turned it into an instrument of privileged propaganda'. French dictionaries have not excluded English borrowings: they continue to include a significant number of Anglicisms, but to cater to a conservative readership, lexicographers produce discourses (in the dictionary paratext and in the media) in which Anglicisms are negatively connoted (Martinez 2011). There is no denying that numerous Anglicisms are unnecessary because equivalents exist in the recipient language. However, as we have seen, the PR criticises Anglicisms that have no substitutes, while it does not label Italianisms that have direct translations in French. This dictionary, claiming to be descriptive (which it mainly is), is also normative and 'patrols the borders', taking on the role of a 'gatekeeper momentarily opening up the bastion to new members', in the words of Mugglestone (2015) – with some new members being more stigmatised than others. This gatekeeper function may be one that many of its users expect the dictionary to perform, as was revealed by a large-scale survey carried out prior to the development of *Usito*,[22] an online dictionary of Quebec French (Cajolet-Laganière 2017). There is nothing wrong with the 'gatekeeper function' so long as the dictionary acknowledges this role, makes it clear to the reader and implements consistent labelling, which the PR does not do. In *Usito* (self-described as normative), a typology of Anglicisms that are criticised in various sources has been designed, and different categories trigger different lexicographic treatments. Even if these treatments are not ideal (Poirier 2015) and even if the prescriptive adjective *critiqué* 'criticised' qualifies only Anglicisms, the normative labelling actually takes into account the entrenchment of a word in the recipient language as well as the existence or lack of an equivalent: *deadline* is criticised because it is presented as a non-standard synonym of *date*

butoir, *date limite*, *échéance*. However, *hamburger*, *hot-dog* and *burger* are not labelled. Their North American English origin is mentioned in the etymology section, as for borrowings from other languages: *bye-bye* is labelled FAM. 'colloquial' but is not marked by a diaintegrative label, just as the Italian *ciao* is. Should the PR wish to acknowledge its (self-assigned) descriptive *and* normative role, it would gain from implementing a consistent labelling strategy based on explicit criteria.

4. Conclusion

In this article, we have presented DiCo, a database that contains the list of modifications that have occurred over time in the macrostructure of four French dictionaries. The record of these modifications, established manually by a pairwise comparison of the successive editions of the same dictionary, makes it possible to describe the evolution of a given dictionary over time and to compare certain characteristics of two distinct dictionaries. Quantitative studies can be conducted directly from the resource. We have also illustrated that the database can be used as a starting point for qualitative studies, for example, to delimit a subset of vocabulary to be further examined. These possibilities are all the more valuable, as the scant information provided by French publishing houses on their dictionaries is unlikely to provide sufficient knowledge about the nature of the dictionaries or a deep understanding of the lexicographic process. In addition to a resource for the benefit of experts carrying out metalexicographical studies, DiCo can be useful for linguists interested in lexicology as well as terminologists. Finally, it could be used in NLP, for example, to build specialised lexicons – based on linguistic labels – such as the recent vocabulary of a technical field, a lexicon of diatopic variation, or a lexicon of dated words. This type of lexicon can be useful for document classification or corpus annotation. Adding external information from a collaborative dictionary to DiCo, such as the inclusion date of words in *Wiktionnaire*, not only makes it possible to compare and contrast the professional dictionaries and the so-called amateur dictionary, but also reveals irregularities and sheds light on specific events occurring in the evolution of the dictionaries under study that deserve further scrutiny. Close examination often reveals that changes occurring in a dictionary, as well as the differences observed between two dictionaries that are supposed to describe the same vocabulary (and its usage) of the same language in the same society, do not always reflect linguistic facts but rather are bound to the lexicographic process and editorial policies.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. We are also grateful to John Humbley for his insightful remarks and suggestions.

Notes

[1] Comments are valid for French lexicography: Canadian lexicography has its own, different, story. For details on the evolution of French lexicography, see Corbin (1998, 2008).

[2] http://redac.univ-tlse2.fr/lexiques/dico.html

[3] Feminine endings and alternative spellings appear in headwords such as *startuper ou startupeur, euse* 'the (male or female) founder of a start-up'. In such cases, headwords are left unchanged.

[4] A total of 47% of the new entries added to the 9th edition of the DAF are technical terms, according to the domain labels.

[5] WIND (Wiktionary INclusion Dates) is a resource developed by Sajous et al. (2020b), that contains the inclusion dates of French and English words in the nomenclature of *Wiktionnaire* and *Wiktionary*.

[6] The typology of Hausmann et al. (1989) is a refinement of the typology devised by Hausmann (1977). It comprises eleven types of labels (as opposed to eight in 1977), to which we added a type called *diasemantic* to signal labels denoting a semantic link between two senses (*par extension* 'by extension', *spécialement* 'especially', *par métonymie* 'by metonymy', etc.). See Corbin and Gasiglia (2017) for details on the recording of linguistic variation in French dictionaries and Ptaszynski (2010) for a survey of research on diasystems.

[7] This choice also prevents the display overload that the eleven types of Hausmann et al.'s typology would cause in the browser presented in Figure 1, by scattering similar information in numerous columns.

[8] Regarding word additions in DiCo, there are only four occurrences of the *populaire* label in the PR (two of which are collocated with *familier* 'colloquial') and none in the PL since (at least) 1998.

[9] http://www.macmillandictionaries.com/features/labels-and-abbreviations/ (accessed on 12 February 2020).

[10] https://public.oed.com/updates/ (accessed on 12 February 2020).

[11] These observations relate only to 'regular' articles.

[12] Boxplots have been generated with R statistical software, with data directly extracted from DiCo as inputs. The first and third quartiles of the distribution are represented by the lower and upper limits of the box, and the median value is represented by the horizontal line inside the box. The lower and upper horizontal lines outside the box represent the minimum and maximum values, excluding possible outliers, and the circles beyond these limits correspond to possible outliers, or extreme values.

[13] Words attested for several centuries that recently entered the PR are mostly regionalisms and words coming from the Francophonie (and, to a lesser extent, technical words). Such words were massively added on the occasion of the 2007 redesign (with PR2007 and PR2008 showing the greatest number of extreme values). For instance, *appondre* 'to attach, to tie' entered PR2007, where it is dated 1165-70 and labelled as an informal regionalism. Other extreme values are due to the (conscious or unconscious) rediscovery of disappeared words. For example, according to the PR, *déceptif* (dated 1378) reappeared in the 20th century under the influence of the English *deceptive* and entered PR2017. The pejorative word *ochlocratie* 'ochlocraty' (mob rule) had not disappeared but was very rare until it started been used by some people to denigrate the yellow vests movement (a grassroots protest movement that began in France in October 2018). Dated 1534, the word entered PR2020.

[14] New editions of the PR and the PL are released by late spring. The inclusion date of their new words was set as 1 June to perform the comparisons between *Wiktionnaire* and the PL/PR.

[15] An online query on the whole PR returns only one word (*statthalter*) preceded by 'GERMANISME', one sense (*restauration #2*, a diatopic variant for *restaurant*) preceded by 'RÉGION. (germanisme)' and one Italianism (*bravoure #2*, a dated term in the musical domain). These indications are undocumented in the PR paratext and are not written with the same typographical conventions as linguistic labels.

[16] For more details on French language policy and authoritative institutions, we refer the reader to Humbley (2008) and Saugera (2017).

[17] http://www.culture.fr/franceterme (accessed on 12 February 2020).

[18] The PR describes *enchilada*, *fajita* and *burrito* as '*mots hispano-américains'* corresponding, according to the PR definition of this adjective, to the Spanish language spoken in Latin America.

[19] *Causette* was the French equivalent for *chat* in 2002, when the word entered the PR. Another official equivalent (*dialogue en ligne*) was adopted in 2006, and the dictionary article was updated accordingly in the 2009 edition.

[20] The French substitute given by the PR for *podcast* is the official equivalent for *podcasting*, published in the *Journal Officiel* dated 15 December 2006. In the *Journal Officiel* dated 23 May 2020, new equivalents of '*podcast* and its derivatives' replace the 2006 substitute: *audio, audio à la demande*, *programme* ou *émission à la demande* should be used to denote the audio file. Other official substitutes newly replace the act of downloading a podcast and the online service.

[21] Among vehement opponents of Anglicisms, Étiemble (1964) mocked this Anglomania in his essay *Parlez-vous franglais ?*.

[22] https://usito.usherbrooke.ca/ (accessed on 12 February 2020).

References

A. Dictionaries

[DAF8-1] Dictionnaire de l'Académie française (1932), Paris, Hachette, huitième édition, tome 1 (A - G).

[DAF8-2] Dictionnaire de l'Académie française (1935), Paris, Hachette, huitième édition, tome 2 (H - Z).

[DAF9-1] Dictionnaire de l'Académie française (rééd. 2001), Paris, Librairie Arthème Fayard / Imprimerie nationale éditions, neuvième édition, tome 1 (A - Enz), XII + 852 + VIII p.

[DAF9-2] Dictionnaire de l'Académie française (2000), Paris, Librairie Arthème Fayard / Imprimerie nationale éditions, neuvième édition, tome 2 (Éoc - Map), VI + 596 + IV p.

[DH2018] Dictionnaire Hachette 2018 (2017), Vanves, Hachette Education, 1872 p.

[PL1906] Petit Larousse illustré 1906 (fac-similé, 2004), Paris, Larousse, 1664 p.

[PL1907] Petit Larousse illustré (millésime : page manquante), Paris, Librairie Larousse, 1664 p.

[PL1908] Petit Larousse illustré 1908 (1907), Paris, Librairie Larousse, 1664 p.

[PL1909] Petit Larousse illustré 1909 (1908), Paris, Librairie Larousse, 1664 p.

[PL1910] Petit Larousse illustré 1910 (1909), Paris, Librairie Larousse, 1664 p.

[PL1911] Petit Larousse illustré 1911 (1910), Paris, Librairie Larousse, 1664 p.

[PL1912] Petit Larousse illustré 1912 (1911), Paris, Librairie Larousse, 1664 p.

[PL1913] Petit Larousse illustré 1913 (1912), Paris, Librairie Larousse, 1664 p.

[PL1914] Petit Larousse illustré 1914 (1913), Paris, Librairie Larousse, 1664 p.

[PL1915] Petit Larousse illustré 1915 (1914), Paris, Librairie Larousse, 1664 p.

[PL1916] Petit Larousse illustré 1916 (1915), Paris, Librairie Larousse, 1664 p.

[PL1917] Petit Larousse illustré 1917 (1916), Paris, Librairie Larousse, 1664 p.

[PL1918] Petit Larousse illustré 1918 (1917), Paris, Librairie Larousse, 1664 p.

[PL1919] Petit Larousse illustré 1919 (1918), Paris, Librairie Larousse, 1664 p.

[PL1920] Petit Larousse illustré 1920 (1919), Paris, Librairie Larousse, 1664 + XVI p.

[PL1921] Petit Larousse illustré 1921 (1920), Paris, Librairie Larousse, 1664 + XVI p.

[PL1922] Petit Larousse illustré 1922 (1921), Paris, Librairie Larousse, 1664 + XVI p.

[PL1923] Petit Larousse illustré 1923 (1922), Paris, Librairie Larousse, 1664 + XVI p.

[PL1924] Petit Larousse illustré 1924 (1923), Paris, Librairie Larousse, 1664 + XVI p.

[PL1925] Petit Larousse illustré 1925 (1924), Paris, Librairie Larousse, 1760 p.

[PL1997] Petit Larousse illustré 1997 (1996), Paris, Larousse, 1784 p.

[PL1998] Petit Larousse illustré grand format 1998 (1997), Paris, Larousse / Bordas, 1870 p.

[PL1999] Petit Larousse illustré 1999 (1998), Paris, Larousse / Bordas, 1784 p.

[PL2000] Petit Larousse illustré grand format 2000 (1999), Paris, Larousse / HER, 1870 + LXXX p.

[PL2001] Petit Larousse illustré 2001 (2000), Paris, Larousse / HER, 1786 + LXXX p.

[PL2002] Petit Larousse illustré grand format 2002 (2001), Paris, Larousse / VUEF, 1852 + XCVI p.

[PL2003] Petit Larousse 2003 (2002), Paris, Larousse / VUEF, 1818 + CXII p.

[PL2004] Petit Larousse illustré 2004 (2003), Paris, Larousse / VUEF, 1818 + CXII p.

[PL2005] Petit Larousse illustré 2005 (2004), Paris, Larousse, 1856 + CXII p.

[PL2006] Petit Larousse illustré 2006 (2005), Paris, Larousse, 1856 + CXXVIII p.
[PL2007] Petit Larousse illustré 2007 (2006), Paris, Larousse, 1856 + XCVI p.
[PL2008] Petit Larousse illustré 2008 (2007), Paris, Larousse, 1812 + CVIII p.
[PL2009] Petit Larousse illustré 2009 (2008), Paris, Larousse, 1812 + CXXIV p.
[PL2010] Petit Larousse illustré 2010 (2009), Paris, Larousse, 1818 + CXXXII p.
[PL2011] Petit Larousse illustré 2011 (2010), Paris, Larousse, 1812 + CXL p.
[PL2012] Petit Larousse illustré 2012 (2011), Paris, Larousse, 1910 + LXXII p.
[PL2013] Petit Larousse illustré 2013 (2012), Paris, Larousse, 1934 + LXXII + 10 p.
[PL2014] Petit Larousse illustré 2014 (2013), Paris, Larousse, 2016 p.
[PL2015] Petit Larousse illustré 2015 (2014), Paris, Larousse, 2048 p.
[PL2016] Petit Larousse illustré 2016 (2015), Paris, Larousse, 2044 p.
[PL2017] Petit Larousse illustré 2017 (2016), Paris, Larousse, 2044 p.
[PL2018] Petit Larousse illustré 2018 (2017), Paris, Larousse, 2044 p.
[PL2019] Petit Larousse illustré 2019 (2018), Paris, Larousse, 2044 p.
[PL2020] Petit Larousse illustré 2020 (2019), Paris, Larousse, 2044 p.
[PR1997] Nouveau Petit Robert grand format (1996), Paris, Dictionnaires Le Robert, XXXVI + 2556 p.
[PR1998] Nouveau Petit Robert (1997), Paris, Dictionnaires Le Robert, XXXVI + 2556 p.
[PR1999] Nouveau Petit Robert (1998), Paris, Dictionnaires Le Robert, XXXVI + 2556 p.
[PR2000] Nouveau Petit Robert (1999), Paris, Dictionnaires Le Robert, XXXVI + 2556 p.
[PR2001] Nouveau Petit Robert (2000), Paris, Dictionnaires Le Robert, XXXVI + 2844 p.
[PR2002] Nouveau Petit Robert (2001), Paris, Dictionnaires Le Robert, XXXVI + 2844 p.
[PR2003] Nouveau Petit Robert (2002), Paris, Dictionnaires Le Robert, XXXVIII + 2954 p.
[PR2004] Nouveau Petit Robert grand format (2003), Paris, Dictionnaires Le Robert, XXXVIII + 2954 p.
[PR2005] Nouveau Petit Robert (2004), Paris, Dictionnaires Le Robert, XXXVIII + 2954 p.
[PR2006] Nouveau Petit Robert 2006 (2005), Paris, Dictionnaires Le Robert, XXXVIII + 2954 p.
[PR2007] Nouveau Petit Robert 2007 (2006), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2008] Nouveau Petit Robert 2008 (2007), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2009] Nouveau Petit Robert 2009 (2008), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2010] Nouveau Petit Robert 2010 (2009), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2011] Petit Robert 2011 (2010), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2012] Petit Robert 2012 (2011), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2013] Petit Robert 2013 (2012), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2014] Petit Robert 2014 (2013), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2015] Petit Robert 2015 (2014), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2016] Petit Robert 2016 (2015), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2017] Petit Robert 2017 (2016), Paris, Dictionnaires Le Robert, XLII + 2838 p.
[PR2018] Petit Robert 2018 (2017), Paris, Dictionnaires Le Robert, XL + 2840 p.
[PR2019] Petit Robert 2019 (2018), Paris, Dictionnaires Le Robert, XLI + 2840 p.
[PR2020] Petit Robert 2020 (2019), Paris, Dictionnaires Le Robert, XL + 2840 p.
[WIKT] Wiktionnaire, the French language edition of Wiktionary. https://fr.wiktionary.org/

B. Other literature

Abecassis, M. 2008. 'The Ideology of the Perfect Dictionary: How Efficient Can a Dictionary Be.' *Lexicos* 18: 1-14.

Baroni, M. and Bernardini, S. 2004. 'Bootcat: Bootstrapping Corpora and Terms from the Web.' In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, May 26-28, 2004*, 1313-1316.

Cajolet-Laganière, H. 2017. 'Le traitement des anglicismes critiqués dans le dictionnaire en ligne Usito' In *Les anglicismes : des emprunts à intérêt variable? Actes du colloque du réseau OPALE, Quebec, Canada, October 18-19, 2016*, 128-150.

Corbin, P. 1998. 'La lexicographie française est-elle en panne ?' In *Cicle de Conferències 96-97, Lèxic, corpus i diccionaris, Barcelona, Spain*, 83-112.

Corbin, P. 2008. 'Quel avenir pour la lexicographie française ?' In *Actes du Congrès Mondial de Linguistique Française (CMLF 2008), Paris, France, July 9-12, 2008*, 1227-1250.

Corbin, P. and Gasiglia, N. 2011. 'Éléments pour un état de la description de la variété des usages lexicaux dans les dictionnaires français monolingues (1980-2008).' In Baider, F., Lamprou, E. and Monville-Burston, M. (eds), *La marque en lexicographie. États présents, voies d'avenir*, Limoges: Lambert-Lucas, 17-37.

Corbin, P. and Gasiglia, N. 2017. 'Un demi-siècle de conceptions du traitement de la variation dans la lexicographie d'expression française.' *Revue de Sémantique et Pragmatique* 41-42:15-39.

Étiemble, R. 1964. *Parlez-vous franglais ?* Paris: Gallimard.

Fiévet, A.-C. and Podhorná-Polická, A. 2011. 'La notion d'« argot commun des jeunes » : approches lexicographique et socio-didactique dans les cours de FLE.' In Bastian, S. and Goudaillier, J.-P. (eds), *Registres de langue et argot(s) : lieux d'émergence, vecteurs de diffusion*. München: Martin Meidenbaue, 371-389.

Hathout, N., Sajous, F. and Calderone, B. 2014. 'GLÀFF, a Large Versatile French Lexicon.' In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, May 26-31, 2014*, 1007-1012.

Hausmann, F. J., Reichmann, O., Wiegand, H. E., and Zgusta, L. 1989. *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie*. New-York: Walter de Gruyter.

Hausmann, F. J. 1977. *Einführung in die Benutzung der neufranzösischen Wörterbücher*. Tübingen: Max Niemeyer Verlag.

Humbley, J. 2008. 'How to Determine the Success of French Language Policy on Anglicisms - Some Methodological Considerations.' In Fischer, R and Pułaczewska, H. (eds), *Anglicisms in Europe: Linguistic Diversity in a Global Context*. Newcastle upon Tyne: Cambridge Scholars Publishing, 85-105.

Koch, P. and Oesterreicher, W. 2001. 'Gesprochene Sprache und geschriebene Sprache.' *Lexikon der Romanistischen Linguistik*, Tübingen: Max Niemeyer Verlag, 584-627.

Landau, S. 2001. *Dictionaries: The Art and Craft of Lexicography.* Cambridge: Cambridge University Press.

Lodge, A. 1989. 'Speaker's perception of non-standard vocabulary in French.' *Zeitschrift für romanische Philologie* 105.5-6: 427-444.

Martinez, C. 2009a. *L'évolution de l'orthographe dans les* Petit Larousse *et les* Petit Robert *1997-2008 : une approche généalogique du texte lexicographique*. Ph.D. Thesis, Université de Cergy-Pontoise.

Martinez, C. 2009b. 'Une base de données des entrées et sorties dans la nomenclature d'un corpus de dictionnaires : présentation et exploitation.' *Études de linguistique appliquée* 156:499-509.

Martinez, C. 2010. 'Mots nouveaux du Petit Robert 2011'. Accessed on 12 February 2020. https://orthogrenoble.net/mots-nouveaux-dictionnaires/entrees-petit-robert-2011/.

Martinez, C. 2011. 'Intégration des emprunts dans les *Petit Larousse* et les *Petit Robert* 1997 à 2009. Évolution des nomenclatures et des graphies.' In Steuckardt, A., Leclerq, O.,

Niklas-Salminen, A., and Thorel, M. (ed.), *Les dictionnaires et l'emprunt (XVIe-XXIe siècles)*. Aix-en-Provence: Presses de l'Université de Provence, 247-261.

Martinez, C. 2013. 'La comparaison de dictionnaires comme méthode d'investigation lexicographique.' *Lexique* 21:193-220.

Mugglestone, L. 2015. 'Description and Prescription in Dictionaries.' In Durkin, P. (ed.), *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, 546-560.

Namatende-Sakwa, L. 2011. 'Problems of Usage Labelling in English Lexicography.' *Lexicos* 21: 305-315.

Podhorná-Polická, A. 2011. 'L'expressivité et la marque lexicographique : étude comparative franco-tchèque d'un corpus du lexique non-standard. Les marques *fam.*, *pop.*, *arg*. vs *expressivité* en lexicographies française et tchèque.' In Baider, F., Lamprou, E., and Monville-Burston, M. (eds), *La marque en lexicographie. États présents, voies d'avenir*. Limoges: Lambert-Lucas, 209-225.

Poirier, C. 2015. 'USITO : un pas en avant, un pas en arrière'. *Cahiers de Lexicologie* 106:21-53.

Ptaszynski, M. O. 2010. 'Theoretical Considerations for the Improvement of Usage Labelling in Dictionaries: A Combined Formal-Functional Approach.' *International Journal of Lexicography* 23.4: 411-442.

Rangel, F., Rosso, P., Potthast, M., and Stein, B. 2017. 'Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter.' In *Working Notes Papers of the CLEF 2017*.

Rey-Debove, J. 1971. *Étude linguistique et sémiotique des dictionnaires français contemporains*. Paris-La Haye: Mouton.

Sajous, F., Hathout, N. and Calderone, B. 2014. 'Ne jetons pas le Wiktionnaire avec l'oripeau du web ! Études et réalisations fondées sur le dictionnaire collaboratif.' In *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014), Berlin, Germany, July 19-23, 2014*, 663-680.

Sajous, F., Josselin-Leray, A., and Hathout, N. 2018. 'The Complementarity of Crowdsourced Dictionaries and Professional Dictionaries viewed through the Filter of Neology.' *Lexis* 12.

Sajous, F., Josselin-Leray, A., and Hathout, N. 2020a. 'Les domaines de spécialité dans les dictionnaires généraux : le lexique de l'informatique analysé par les foules et par les professionnels... de la lexicographie.' *Neologica* 14: 83-107.

Sajous, F., Calderone, B. and Hathout, N. 2020b. 'ENGLAWI: From Human- to Machine-Readable Wiktionary.' In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), Marseille, France, May 11-16, 2020*, 3016-3026.

Saugera, V. 2017. *Remade in France: Anglicisms in the Lexicon and Morphology of French*. Oxford: Oxford University Press.

Tiedemann, J. and Ljubešić, N. 2012. 'Efficient Discrimination Between Closely Related Languages.' In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), December 8-15, 2012, Mumbai, India*, 2619-2634.

Vrbinc, M., and Vrbinc, A. 2017. 'Multiple Labels Marking Connotative Values of Idioms in the Oxford Idioms Dictionary for Learners of English.' *3L: The Southeast Asian Journal of English Language Studies* 23.2:96-108.

Wild, K. 2008. 'Vulgar and Popular in Johnson, Webster and the OED.' In *Proceedings of the 13th EURALEX International Congress, July 15-19, 2008, Barcelona, Spain*, 1209-1214.