RiCL Research in Corpus Linguistics

How to build a corpus for a tool-based approach to determinologisation in the field of particle physics

Julie Humbert-Droz - Aurélie Picton - Anne Condamines
University of Geneva / Switzerland & University of Toulouse 2 / France
University of Geneva / Switzerland
CNRS & University of Toulouse 2 / France

Abstract – This paper discusses corpus design and building issues when dealing with a complex, multidimensional phenomenon such as determinologisation. Its representation in corpus data imposes an original reflection on the process and on some essential concepts of corpus building. This paper focuses on the necessity of representing the progressive aspects of determinologisation in the corpus, i.e. through levels of specialisation and through time, and the practical issues this raises. At the same time, it shows that a representative corpus of determinologisation in a specific domain (in this case, particle physics) implies clear and objective criteria when it comes to picking individual texts. Four principles are established to this end. The discussion leads to the proposal of a solid text selection procedure, which ensures that the peculiarities of determinologisation in the domain of particle physics are reflected in the corpus.

Keywords – corpus-building; determinologisation; comparable corpora; tool-based approach; representativeness; textual terminology

1. INTRODUCTION

The increasing use of corpora in linguistics and terminology has made it possible to address various research topics, such as terminological variation (e.g. Freixa 2006; Fernández-Silva 2016; Drouin *et al.* 2017), term and relation identification (e.g. Drouin 2003; León-Araúz *et al.* 2016; Daille 2017), circulation of terms outside of experts' sphere (e.g. Ungureanu 2006; Nicolae and Delavigne 2013; Condamines and Picton 2014). Considering the importance of corpora, numerous research papers and textbooks continually address typical issues related to corpus design (e.g. Biber 1993; Meyer and Mackintosh 1996; Kennedy 1998; Pearson 1998; Habert 2000; Ahmad and Rogers 2001; Bowker and Pearson 2002; McEnery and Hardie 2012). In this view, it is usually

Research in Corpus Linguistics 7: 1–17 (2019). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.07.01

argued that the design of a corpus needs to be thought out in accordance with the purpose for which the corpus is built. This means that the material included in the corpus must reflect the complexity of the phenomenon investigated.

In this context, this paper aims to discuss corpus design and building issues when dealing with a complex, multidimensional linguistic phenomenon such as determinologisation, studied from the viewpoint of one domain (particle physics), in French. We argue that the specificities of determinologisation impose an original and renewed reflection about corpus design, especially with regard to the issue of representativeness. The discussion leads to the building of one corpus, which will be referred to as *PPC* (*Particle Physics Corpus*).

This paper is structured as follows: Section 2 gives a brief overview of the background of the study for which the PPC is built. In Section 3, we attempt to operationalise the concept of representativeness and present the principles that were developed to this end. Section 4 outlines some concluding remarks and states the challenges that lie ahead for the exploration of the corpus.

2. BACKGROUND

2.1. Analysing determinologisation

Determinologisation can be understood both as the process by which terms move from specialised language (LSP) into everyday language and as its result, i.e. the use of terms in a non-specialised context (Guilbert 1975; Meyer and Mackintosh 2000; Ungureanu 2006). In the latter case, it is known that semantic changes are likely to occur, such as the appearance of a shallower meaning, or metaphors, or word play (Meyer and Mackintosh 2000; Renouf 2017). As for the process, it can be considered as continuous on two levels. First, terms probably do not move into non-specialised language directly. Rather, they might be used in different genres and different levels of LSP communication in the process (Halskov 2005; Condamines and Picton 2014). Second, terms progressively integrate general language over time (Dury 2008; Renouf 2017).

In this context, our research aims to gain a broader understanding of determinologisation as a continuous process, one aspect that has received less attention than the use of terms in non-specialised texts.[1] Our purpose is twofold: on the one hand, we aim to identify different factors that cause a term to determinologise, and on the other hand, we seek to better understand the role of the various media through which terms reach general language. In fact, the PPC is the first step of this research project. Its design results from a thorough analysis of the way in which determinologisation works, the diversity of texts involved in the process, and their representation in textual data.

2.2. A textual terminology methodology

This approach is based on the principles of Textual Terminology (e.g. Bourigault and Slodzian 1999; Condamines 2003), in which the analysis is usually conducted on texts and in collaboration with domain experts. Such importance is given to textual data because "it is in the texts produced or used by a community of experts that most of the knowledge shared by this community is expressed, and thus accessible. Therefore, the analysis must begin there"[2] (Bourigault and Slodzian 1999: 30). From this viewpoint, specialised texts, usually gathered in corpora, constitute the primary material on which linguistic analyses are carried out, mostly from a tool-based approach. This approach mainly relies on comparable corpora, in which linguistic clues that are associated to the phenomena under study are explored, and the organisation of the data in sub-corpora is determined by the research purpose.

The differences that emerge from sub-corpora comparisons are interpreted in relation to the research purpose, and with the help of domain experts. In reality, since the analyst is usually not an expert of the domain under study, domain experts contribute to the analysis from the corpus compilation to the interpretation and validation of results. The whole analysis is thus a 'co-construction' process (Picton 2011: 137).

2.3. Particle physics as a 'determinologisable' domain

In order to conduct a systematic analysis of determinologisation in a domain, one must first ensure that terms from this domain are likely to integrate general language.

[1] See Meyer and Mackintosh (2000), Ungureanu (2006), Renouf (2017) for such studies.

[2] Our translation.

Following Guilbert (1975: 84), we assume that such terms belong to domains that regularly appear in the media. Many domains could satisfy this condition, but we believe that particle physics is particularly relevant given the rather recent mediatisation of the building and exploitation of the LHC (Large Hadron Collider) as well as the Higgs boson discovery.

3. Representing determinologisation in corpus data: Four essential principles

In the previous section, we gave a brief overview of the study for which this corpus is built as well as the theoretical and methodological context. We will now explain how the concept of representativeness can be operationalised through a solid compilation method.

It must be pointed out, though, that because this concept has been extensively discussed in numerous research papers,[3] we will not further debate it. Rather, our point is that representativeness is an ideal that should be approached. To do so, we developed a strategic and informed decision-making process, which deals with the necessary heterogeneity of the data. Indeed, representing determinologisation in corpus data implies the inclusion of texts from different levels of specialisation, different genres and different time periods. In addition, our research project being restricted to a specific domain, the presence of relevant terms in the corpus must be ensured. This is achieved through a compilation method that relies on four principles. At this point, let us underline that, although this method is applied to French, it is language independent and can therefore be adapted to any other language. Some choices may differ, especially when it comes to identifying relevant genres, but the basic principles described in this paper remain valid.

3.1. From highly specialised language to general language

The first principle relates to the level of specialisation of the texts. All the levels involved in the determinologisation process are to be considered, from highly specialised to general language, and the criteria established to determine these levels are discussed here.

[3] For example, Sinclair (1991), Biber (1993), Kennedy (1998), Habert (2000), Leech (2007).

First, it seems obvious that highly specialised language and general language should be included. The former is often represented by texts from what Bowker and Pearson (2002: 28) call an expert-expert level of LSP communication but the latter is a more complex concept (e.g. Ahmad and Rogers 2001: 735). Besides, because representing general language in a corpus is even more complex,[4] newspaper corpora are often considered to be an adequate operational choice (e.g. Halskov 2005; Dury 2008; Renouf 2017).

Second, let us consider the other two levels of LSP communication identified by Bowker and Pearson (2002: 28), which seem to be the most relevant to describe a communication level that is 'in between' (not highly specialised and not general). These are from experts to semi-experts and from experts to non-experts. The difference between semi-experts and non-experts mainly relies on people's knowledge of a subject. Semi-experts are considered to have some knowledge of a domain, whereas non-experts are considered to have none (or almost none). In reality, the difference might be subtler and more difficult to assert because it depends on the knowledge each individual has of a subject. Considering that the readership of a text might be very diverse, it seems even more difficult to determine the level of (non-)expertise of each individual. Therefore, although this distinction between semi- and non-experts is clear in theory, it does not seem fully operational here. This is why we would rather not distinguish between these two and include them both in the term 'intermediate level of specialisation'.

Texts from these three different levels of specialisation are essential to represent one progressive aspect of determinologisation and at least three sub-corpora should compose the corpus (one for each level identified). Let us now examine which genres are likely to best represent this process for each of these levels.

3.2. The importance of text genres

According to Bhatia (2004: 23), "genre essentially refers to language use in a conventionalized communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms by imposing constraints on the use of lexico-grammatical as well as

[4] According to the large number of criteria extensively discussed for general language corpora (e.g. Sinclair 1991; Kennedy 1998; Siepmann *et al.* 2017).

discoursal resources.” Many genres are found in LSP, which can be rather diverse according to both the level of specialisation and the domain (Meyer and Mackintosh 1996: 270). In our case, though, only the genres that are relevant for our research purposes must be selected, despite their diversity. Their identification relies on three main principles: 1) the genres must be likely to participate in the transfer of terms into general language, 2) they must be consistent with the levels of specialisation identified, and 3) they must be relevant for the domain under investigation.

In the following, and for explanatory purposes, we will first discuss the genres that compose the specialised and the non-specialised parts of the PPC and then those that are included in the intermediate part of the PPC.

First, according to Loffler-Laurian (1983: 10*sqq.*) and Bowker and Pearson (2002: 28), specialised articles are particularly relevant to represent highly specialised language. However, since there is a lack of this type of publication in French, doctoral theses were also considered.

Second, following the majority of authors who studied determinologisation from a corpus linguistics viewpoint, we compiled a corpus of general newspaper articles as a way to represent non-specialised language. In addition to the practical reasons discussed in 3.1, this choice implies a deeper conceptual motivation. On the one hand, Meyer and Mackintosh (2000: 112) argue that determinologisation describes the phenomenon that occurs “when a term captures the interest of the general public.” It is assumed that this interest is reflected in the topics addressed by the media, hence in the terms they use. On the other hand, it is well known that the media are of great influence in this process (e.g. Cabré 1994: 593; Pearson 1998: 26; Moirand 2007: 20).

Third, the identification of genres that are relevant for the intermediate part of the corpus requires two additional conditions, which are complementary to the principles stated above. They are mostly based on two studies in which such texts are exploited: Condamines and Picton (2014: 171*sqq.*), who compiled a corpus of press releases, and Halskov (2005: 54*sqq.*), who used a corpus composed of science popularisation articles and ‘newsgroup postings’, among other genres. According to these studies, various genres from an intermediate level of specialisation are likely to play a part in the transfer of terms into general language, either directly or in a more indirect way. In this context, we believe that different genres should be included in order to best represent this diversity.

Therefore, the first condition is that anyone must be able to find and read the texts; they must not be restricted to a certain community. This will be referred to as the 'availability' condition. The second condition is based on the concept of 'knowledge transfer discourse'[5] (Beacco and Moirand 1995). According to the authors, many genres participate in transferring knowledge, even when it is not their primary purpose. Moreover, since knowledge is usually transferred through terms, terms probably appear in these genres, making them particularly relevant for our study. Thus, any genre that is described as a type of knowledge transfer discourse is considered as relevant for the intermediate part of the corpus. These conditions allow us to disregard genres such as textbooks, which seem to be restricted to a rather well delimited speech community. Genres such as press releases, general reports and science popularisation websites and articles appear to be much more adequate, as we explain below.

First of all, since press releases are by definition intended for journalists (Nicolae and Delavigne 2013: 219), journalists are likely to reuse the terms they find in press releases (Condamines and Picton 2014: 171*sqq.*). In this view, they represent a step in the transfer of terms from LSP to general language, more precisely from experts to journalists, and are thus relevant.

Secondly, we included general reports from several particle physics research laboratories. Annual (or biennial) general reports aim to inform the public about research activities. As such, they are considered as a type of knowledge transfer discourse. Moreover, since these reports are usually freely available online and may be read by anyone, the availability condition is also satisfied.

Lastly, we took into consideration two science popularisation genres. According to Guilbert (1975), for example, science popularisation and determinologisation are two closely related concepts, though the link between them is not clear. Authors such as Jacobi (1986) or Delavigne (2001) argue that science popularisation is an intermediary between experts and non-experts, with its main purpose being to transmit knowledge (Delavigne 2001: 28). As such, popularisation genres are particularly relevant. Moreover, to better represent the diversity of science popularisation media, we included two complementary genres: articles and websites. Indeed, journal articles are likely to treat current topics, such as news or discoveries, whereas websites tend to explain a domain in a more general way.

[5] Our translation.

To sum up, considering the arguments that were advanced in this section, the specialised part of the corpus is composed of specialised articles and theses, the intermediate part includes press releases, general reports, popularisation articles and websites, and the non-specialised part contains general newspaper articles.

3.3. Representing both dimensions of progression through a double division into sub-corpora

The main challenge of this corpus design lies in finding a way to represent both progressive aspects of determinologisation, through levels of specialisation and through time. Given that our approach relies on a comparable corpus, this third principle is about the organisation of the data in sub-corpora. More precisely, we argue that two types of sub-corpora are necessary to reflect both dimensions of progression.

3.3.1. First dimension of progression: Through levels of specialisation

As we mentioned in Section 3.1, the corpus should consist of at least three sub-corpora, one for each level of specialisation. However, in Section 3.2 we identified different genres that are relevant for these levels and for our research purposes. Therefore, based on the assumption that terms might behave in specific ways according to the genre, considering these genres separately seems more relevant. As a matter of fact, at this point, we do not know if the behaviours that we assume we will observe are related to the genre in which the terms are used, or to determinologisation – or both. Nevertheless, some genres should be grouped together, either because of their relative similarity and complementarity (popularisation articles and websites) or because of the practical reasons explained in 3.2 (specialised articles and theses).

Sub-corpus	**Level of specialisation**	**Text genre**
Specialised	High	Specialised journal articles Doctoral theses
Press releases (PR)	Intermediate	Press releases
Reports	Intermediate	Laboratory general reports
Science popularisation (SPop)	Intermediate	Journal articles Websites
Press	Non-specialised	General newspaper articles

Table 1: Composition of the PPC

As a result, the PPC is composed of five sub-corpora, which represent a way of approaching one continuous aspect of determinologisation, while remaining manageable in relevant corpus analysis tools[6] (see Table 1).

3.3.2. Second dimension of progression: Through time

The second dimension of progression is the diachronic dimension. For the data to reflect it, each of the five sub-corpora discussed above is further divided into smaller sub-corpora. The period to take into account and its division into shorter periods are discussed in this section.

According to Dury and Picton (2009: 38), when investigating evolution in recent or recently changing domains, it is probably more interesting to consider shorter periods, mainly because change can occur rather quickly. Picton (2011) calls such approach *short-term diachrony*. In order to determine these periods, two strategies are usually employed: the division is either arbitrary (e.g. several periods made up of the same number of years) or based on extra-linguistic criteria (Picton 2018: 44). In this case, we mainly rely on extra-linguistic criteria, which are related to the role of the media in determinologisation, and to the assumption that some important events of the domain might influence the ways in which terms are used in the corpus. Moreover, we assume that the media extensively covered these events, thus contributing to the transfer of terms in general language.

In accordance with the principles of Textual Terminology (Section 2.2), we collaborated with an expert to identify two events: the start of the LHC in 2008 and the discovery of the Higgs boson in 2012. Consequently, the corpus covers the period from 2003 to 2016 (2016 being the compilation time) and is organised in three shorter periods: 1) from 2003 to 2007, 2) from 2008 to 2011, and 3) from 2012 to 2016. Texts published prior to 2003 were not included so that the sub-corpora remain balanced and comparable. The corpus is thus composed of fifteen sub-corpora, as shown in Table 2.

[6] Indeed, it seems almost impossible to handle too many sub-corpora with current tools.

	Progression through time		
Progression through levels of specialisation	Specialised-2003-2007	Specialised-2008-2011	Specialised-2012-2016
	PR-2003-2007	PR-2008-2011	PR-2012-2016
	Reports-2003-2007	Reports-2008-2011	Reports-2012-2016
	SPop-2003-2007	SPop-2008-2011	SPop-2012-2016
	Press-2003-2007	Press-2008-2011	Press-2012-2016

Table 2: Composition of the corpus in terms of sub-corpora

3.4. Ensuring domain relevance through an objective text selection procedure

This fourth principle addresses the issue of how each individual text is selected, which is a more operational viewpoint on corpus compilation. Indeed, not only should the texts be relevant for the determinologisation process (in terms of levels of specialisation, genres and publication dates), but they must be relevant for the domain as well. To this end, we detail a solid text selection procedure for the sub-corpora to remain balanced in terms of content. Paradoxically, this procedure must be flexible enough so that it can be adapted to the necessary heterogeneity of the documents.

For explanatory purposes, we first discuss the sub-corpora containing texts from either a high or intermediate level of specialisation. Second, we argue that this procedure must be refined for the *Press* sub-corpus.

3.4.1. A balance between keywords and experts

One common method of assessing the relevance of texts when building a corpus is based on a usually quick evaluation of their content. According to Pearson (1998: 54), this may be achieved “by looking at what a particular text is about (e.g. on the basis of its title, table of contents in the case of a book)” and by “examining the lexical structure of a text and identifying keywords used frequently in the text.” To this end, we developed an approach based on specific terms considered as key to the domain. Collaborating with an expert was necessary to define a sufficient number of keywords (almost) unequivocally referring to the domain, and in particular to the subdomain of the Standard Model of particle physics.[7] She pointed at terms such as *Modèle Standard*, *boson de Higgs*, *ATLAS*, *LHC* or *particule élémentaire*,[8] and the texts containing them in titles, tables of contents or even sometimes in the body were retained.

In fact, the expert played a determining role throughout this corpus building process. Not only did she identify the relevant events used for the diachronic division, but she also clarified the complex links between CERN (European Organization for Nuclear Research), the Standard Model, the LHC and the Higgs boson, leading us to find the most relevant sources. Therefore, given that the aforementioned events both happened at CERN and that CERN is located at the Swiss-French border, only French and Swiss sources related to this organisation were included. Thus, all the texts come from:

- Swiss and French universities that provide access to theses in French,
- the only French research journal that publishes articles in French,
- Swiss and French websites of laboratories undertaking research in particle physics (including CERN),
- French science popularisation journals, and
- Swiss and French newspapers.

Based on these sources, the overall text selection procedure broadly consisted of 1) listing the individual texts containing at least one of the keywords, 2) discussing and refining the list with the expert, and 3) balancing the size of the diachronic sub-corpora so that they remained comparable. In other words, domain relevance is ensured both by the presence of certain keywords in the texts and by a close collaboration with a domain expert. However, this selection procedure is not adequate for the *Press* sub-corpus and it must be adapted, as we explain in the next section.

3.4.2. Refining the procedure for the particular case of the *Press* sub-corpus

Although our corpus design includes newspaper articles containing particle physics terms, our research purposes require more varied material. Since determinologised terms can behave in various ways in a non-specialised context, the possibility of finding such behaviours in the corpus must be ensured. To do so, however, the text selection

[7] Narrowing down the domain proved necessary given the large number of subdomains comprised in particle physics.

[8] *Standard Model, Higgs boson, ATLAS, LHC, elementary particle.*

procedure cannot rely on the same limited number of keywords as the other four sub-corpora. Otherwise, the articles would be very similar in content. More importantly, the different behaviours resulting from determinologisation would likely be missing.

That being said, a random selection does not seem operational either. Indeed, for particle physics terms to be analysed, we need to make sure that they appear in the corpus – and that they appear frequently enough. Thus, we propose a hybrid method that guarantees the presence of relevant terms, while ensuring that some occurrences are examples of determinologised terms. Articles are selected based on a large number of terms attested in the other four sub-corpora, which are used as keywords on the platform LexisNexis®.[9] Such a large number of keywords provides more diverse results and avoids the bias of selecting rather similar articles. Furthermore, a large number of keywords also maximises the chances of observing a whole variety of contexts, some of which might be linked to determinologisation. But the selection process must be carried out carefully, by choosing not only articles containing many keywords, but also, and more importantly, articles containing few. Indeed, if only one term appears in an article, for example, this one occurrence could be a metaphor or word play, or even another consequence of determinologisation that we do not know of yet.

In this context, an objective method was developed in order to identify the terms that are attested in the other sub-corpora and that are relevant to retrieve the articles on LexisNexis®. It is illustrated in Figure 1.

[9] It is accessed via a subscription at the University of Geneva.

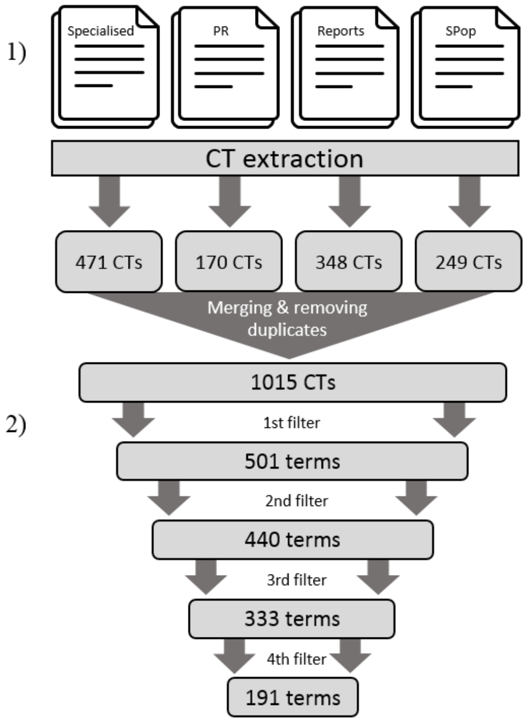

Figure 1: Term extraction and keyword selection

The method broadly consists of 1) a candidate term (CT) extraction and 2) a refinement of the list of CTs through objective filters. The CTs were extracted from the *Specialised*, *PR*, *Reports* and *SPop* sub-corpora with TermoStat (Drouin 2003) and only those with a specificity score higher or equal to 40 were retained. After removing the duplicates, the list consists of 1015 CTs.

As TermoStat is a hybrid term extraction system (Drouin 2003: 99), some of the extracted CTs were in fact noise, whereas other CTs seemed less relevant to build the *Press* sub-corpus. Thus, we applied four filters to keep the most relevant terms only and we removed the following:

a. **CTs that are actually not French terms**. They fall into three categories:

   - proper nouns (e.g. *Perrine, Rolf*), countries (e.g. *République Slovaque*), abbreviations appearing in bibliographical references (e.g. *phys*, *nucl*), parts of URLs or email addresses (e.g. *lhc-france*, *grey@cern*);
   - incomplete CTs, probably due to tagging inaccuracy (e.g. *détecteur de pied*, instead of *détecteur de pied de gerbe*); and

- English terms that are only used in the English parts of the texts, such as references (e.g. *calorimeter, polarization*).

b. **CTs that do not designate domain-specific concepts and that are considered as transdisciplinary** (such as *analyse, interaction, fonctionnement*). This filter is based on the *transdisciplinary scientific lexicon* (e.g. Tutin 2007; Drouin 2007), and we exploited the list established by the Scientext project.[10]

c. **CTs that belong to more than five domains in more than three terminological data banks** (Grand dictionnaire terminologique, Termium, FranceTerme, IATE and TERMDAT), such as *accélération, grille, collision, vitesse*. This step is based on the idea that, although these terms do belong to particle physics, they are likely to be polysemous in newspaper articles. As a result, some of their occurrences may neither designate a concept of particle physics nor convey a determinologised meaning of a particle physics term. Thus, they appear to be inadequate for the selection procedure discussed here.

d. **CTs that are not attested in all of the four sub-corpora**, such as *superchamp, préon, leptoquark.*

The final list is composed of 191 terms. We believe that this procedure allowed us to build a sub-corpus that is relevant for the domain – given that the articles contain one or more of these terms – and that is adequate to explore the consequences of determinologisation – given that they were anticipated through a balanced text selection. This last sub-corpus was the final step of the building process discussed in this paper and it completes the whole corpus (see Table 3).

Sub-corpus	**2003-2007**	**2008-2011**	**2012-2016**	**Total**
Specialised	314,658	330,975	349,242	**994,875**
PR	70,950	69,478	69,892	**210,320**
Reports	516,820	302,552	322,501	**1,141,873**
SPop	216,969	194,675	208,401	**620,045**
Press	367,378	365,650	365,680	**1,098,708**
Total	1,486,775	1,263,330	1,315,716	**4,065,821**

Table 3: Size of the corpus in number of occurrences

[10] Scientext project, https://scientext.hypotheses.org/, last access: 28 April 2019.

4. CONCLUDING REMARKS

In this paper, we presented an original reflection on the design of a corpus meant to be representative of the determinologisation process in particle physics. In particular, we discussed some essential issues regarding corpus building and the specific ways they must be addressed given the progressive dimensions involved in this process. From this viewpoint, we mainly discussed how the concept of 'representativeness' can be operationalised through an objective compilation method that relies on four principles:

1. the texts included in the corpus represent the levels of specialisation involved in the determinologisation process (highly specialised, intermediate, non-specialised);
2. they belong to genres that are likely to take part in this process. This feature was mainly identified based on the concepts of 'availability' and 'knowledge transfer';
3. the progressive aspects of determinologisation are represented by two types of sub-corpora. A division into five sub-corpora reflects progression through levels of specialisation, and each of these sub-corpora is divided into three diachronic sub-corpora, which represent progression through time;
4. the texts are relevant given the domain investigated. A solid and objective text selection procedure was developed to this end.

Our work now focuses on the proper exploration of the corpus. Indeed, such a large number of sub-corpora remains a challenge for any corpus analysis tool and for the analyst. If it is possible to take into account both dimensions involved in determinologisation separately, analysing them simultaneously, as well as their interactions, seems much more challenging. Finding methods that enable us to handle so many sub-corpora is therefore crucial in order to better understand determinologisation.

REFERENCES

Ahmad, Khurshid and Margaret Rogers. 2001. Corpus linguistics and terminology extraction. In Sue E. Wright and Gerhard Budin eds. *Handbook of Terminology Management*. Amsterdam: John Benjamins, 725–760.

Bhatia, Vijay K. 2004. *Worlds of Written Discourses: A Genre-based View*. London: Continuum.

Beacco, Jean-Claude and Sophie Moirand. 1995. Autour des discours de transmission des connaissances. *Langages* 117: 32–53.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

Bourigault, Didier and Monique Slodzian. 1999. Pour une terminologie textuelle. *Terminologies Nouvelles* 19: 19–32.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language. A Practical Guide to Using Corpora*. London: Routledge.

Cabré, M. Teresa. 1994. Terminologie et dictionnaires. *META* 39/4: 589–597.

Condamines, Anne. 2003. *Sémantique et Corpus Spécialisés: Constitution de Bases de Connaissances Terminologiques*. Toulouse: Université Toulouse le Mirail.

Condamines, Anne and Aurélie Picton. 2014. Des communiqués de presse du Cnes à la presse généraliste. Vers un observatoire de la diffusion des termes. In Pascaline Dury, José Carlos de Hoyos, Julie Makri-Morel, François Maniez, Vincent Renner and María Belén Villar Diaz eds. *La Néologie en Langue de Spécialité: Détection, Implantation et Circulation des Nouveaux Termes*. Lyon: Centre de Recherche en Terminologie et Traduction, Université Lumière Lyon 2, 165–188.

Daille, Béatrice. 2017. *Term Variation in Specialised Corpora*. Amsterdam: John Benjamins.

Delavigne, Valérie. 2001. *Les Mots du Nucléaire. Contribution Socioterminologique à une Analyse des Discours de Vulgarisation*. Université de Rouen dissertation.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9/1: 99–117.

Drouin, Patrick. 2007. Identification automatique du lexique scientifique transdisciplinaire. *Revue Française de Linguistique Appliquée* 12/2: 45–64.

Drouin, Patrick, Aline Francoeur, John Humbley and Aurélie Picton eds. 2017. *Multiple Perspectives on Terminological Variation*. Amsterdam: John Benjamins.

Dury, Pascaline. 2008. The rise of carbon neutral and compensation carbone: A diachronic investigation into the migration of vocabulary from the language of ecology to newspaper language and vice versa. *Terminology* 14/2: 230–248.

Dury, Pascaline and Aurélie Picton. 2009. Terminologie et diachronie: Vers une réconciliation théorique et méthodologique? *Revue Française de Linguistique Appliquée* 14/2: 31–41.

Fernández-Silva, Sabela. 2016. The cognitive and rhetorical role of term variation and its contribution to knowledge construction in research articles. *Terminology* 22/1: 52–79.

Freixa, Judit. 2006. Causes of denominative variation in terminology. A typology proposal. *Terminology* 12/1: 51–77.

Guilbert, Louis. 1975. *La Créativité Lexicale*. Paris: Larousse.

Habert, Benoît. 2000. Des corpus représentatifs: De quoi, pour quoi, comment? In Mireille Bilger ed. *Linguistique sur Corpus: Études et Réflexions*. Perpignan: Les Presses de l'Université de Perpignan, 11–58.

Halskov, Jakob. 2005. Probing the properties of determinologization: The DiaSketch. *Lambda* 29: 39–63.

Jacobi, Daniel. 1986. *Diffusion et Vulgarisation: Itinéraires du Texte Scientifique*. Paris: Les Belles Lettres.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Leech, Geoffrey. 2007. New resources, or just better old ones? The Holy Grail of representativeness. In Marianne Hundt, Nadja Nesselhauf and Carolin Biewer eds. *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 133–149.

León-Araúz, Pilar, Antonio San Martín and Pamela Faber. 2016. Pattern-based word sketches for the extraction of semantic relations. In Patrick Drouin, Natalia

Grabar, Thierry Hamon, Kyo Kageura and Koichi Takenchi eds. *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*. Osaka, Japan, 73–82.

Loffler-Laurian, Anne-Marie. 1983. Typologie des discours scientifiques: Deux approches. *Études de Linguistique Appliquée* 51: 8–20.

McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Meyer, Ingrid and Kristen Mackintosh. 1996. The corpus from a terminographer's viewpoint. *International Journal of Corpus Linguistics* 1/2: 257–285.

Meyer, Ingrid and Kristen Mackintosh. 2000. When terms move into our everyday lives: An overview of de-terminologization. *Terminology* 6/1: 111–138.

Moirand, Sophie. 2007. *Les Discours de la Presse Quotidienne. Observer, Analyser, Comprendre*. Paris: Presses universitaires de France, Linguistique nouvelle.

Nicolae, Cristina and Valérie Delavigne. 2013. In Geoffrey Williams ed. *Actes des Sixièmes Journées de la Linguistique de Corpus*. Lorient: Université de Bretagne-Sud, 217–229.

Pearson, Jennifer. 1998. *Terms in Context*. Amsterdam: John Benjamins.

Picton, Aurélie. 2011. Picturing short-period diachronic phenomena in specialised corpora. A textual terminology description of the dynamics of knowledge in space technologies. *Terminology* 17/1: 134–156.

Picton, Aurélie. 2018. Terminologie outillée et diachronie: Éléments de réflexion autour d'une réconciliation. *ASp* 74: 27–52.

Renouf, Antoinette. 2017. Some corpus-based observations on determinologisation. *Neologica* 11: 21–48.

Siepmann, Dirk, Christoph Bürgel and Sascha Diwersy. 2017. The *Corpus de Référence du Français Contemporain* (CRFC) as the first genre-diverse mega-corpus of French. *International Journal of Lexicography* 30/1: 63–84.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tutin, Agnès. 2007. Autour du lexique et de la phraséologie des écrits scientifiques. *Revue Française de Linguistique Appliquée* 12/2: 5–14.

Ungureanu, Ludmila. 2006. *L'Interpénétration Langue Générale-Langue Spécialisée dans le Discours d'Internet*. Paris: Connaissances et Savoirs.

Corresponding author
Julie Humbert-Droz
UNI MAIL
40, Bd du Pont-d'Arve
1211 Genève 4
Switzerland
e-mail: julie.humbert-droz@unige.ch

received: May 2019
accepted: September 2019