Textual Terminology : Origins, Principles and New Challenges

Anne Condamines, Aurélie Picton

Textual Terminology: Origins, Principles and New Challenges

Anne Condamines[1], Aurélie Picton[2]

[1]CLLE, CNRS et Université Toulouse Jean-Jaurès (UMR5263)

[2]TIM/FTI, Université de Genève

anne.condamines@univ-tlse2.fr, *aurelie.picton@unige.ch*

Abstract

This chapter deals with Textual Terminology, an approach that emerged in the early 1990s along with other approaches presented in this book. One of its main characteristics is that it arose from an encounter between related disciplines that focused on the use of text mining (Natural Language Processing) and modeling tools (Knowledge Engineering and Artificial Intelligence). These disciplines have worked together to consider the new needs related to terminology, and their evolution, the tools and the diversification of needs have resulted in significant methodological developments in Textual Terminology. This chapter traces these developments and discuss the original bases, principles, challenges and perspectives that constitute Textual Terminology.

Keywords: Textual Terminology, specialised corpora, methodological principles, challenges, distributional analysis, knowledge-rich contexts, applied needs

1. Introduction

This chapter is a reflection on Textual Terminology from its emergence in the 1990s to the new issues it faces today. The second section of the chapter identifies the main elements that contributed to its foundation in the 1990s. Special importance is placed on the tools and

collaboration with research in Natural Language Processing (NLP), Knowledge Engineering (KE) and Artificial Intelligence (AI). The third part focuses on the methodology of Textual Terminology as well as the systematic framework for tool-assisted methods aimed at analysing specialised corpora. In particular, we discuss the place of users in the overall analysis and stress the need for collaboration at each stage of the process. The fourth part highlights the modalities and possibilities that Textual Terminology should encompass, in the light of the parallel evolution of corpora, needs and approaches in Textual Terminology, NLP, and KE/AI. This perspective enables us to better situate Textual Terminology in the context of current approaches and to propose research issues to be addressed in the short and medium term.

2. The origins of Textual Terminology

Textual Terminology arose from a combination of the following: (i) new needs in relation to terminology; (ii) the development of corpus linguistics; and (iii) the encounter between Natural Language Processing and Knowledge Engineering/Artificial Intelligence.

2.1. New needs in relation to terminology

Until the 1990s, terminology studies were mainly related to translation needs. However, because of the need to manage and analyse digital documents and resources with NLP tools, terminology came into the spotlight as a resource for managing electronic documentation. Tasks such as document storage, information search and retrieval, and even translation all required up-to-date terminologies adapted to real-world use.

The structure of databases had to be adapted to new tools, and their entries also needed optimisation. For example, in a study conducted in 1992 at Matra Marconi Space (an aerospace company, now Airbus Defence and Space), only 15% of the terms found in the

documentation in French concerning test/simulation benches were recorded in the official terminology in the "Vocabulaire des sciences et techniques spatiales"[1].

The need was then no longer limited to the creation of terminology standards. What was required was the creation of repositories that corresponded to actual term usage and were usable by tools. It was thus necessary to base the structure of terminology repositories on the analysis of digital corpora that were beginning to be available in large quantities in companies.

However, in the early 1990s, most terminology studies were carried out within the framework of the General Theory of Terminology (GTT), developed in the 1930s by Eugen Wüster (see chapters by Candel and Humbley in this volume). Wüster was a Viennese engineer, who advocated a prescriptive view of terminology with the objective of guaranteeing non-ambiguous communication in science and technology and especially within companies, both in the same language as well as different ones. Hence, one of the cornerstones of the GTT was biunivocity: one term <> one concept. Polysemy was thus not acceptable. Although Wüster was familiar with the work of linguists such as Saussure, he believed that it was necessary to separate terminology from discursive variations. One explanation for this is that Wüster was firmly convinced that concepts were the starting point for terminology (Wüster [1979] 1985).

The strictly onomasiological approach of the GTT was harshly criticized by corpus linguists, especially when electronic corpora (internet and intranet among others) became widely available.

2.2. Corpus linguistics and corpus-based Terminology

[1] Edited by the DGLFLF (Délégation Générale à la Langue Française et aux Langues de France, General Delegation for the French Language) https://www.culture.gouv.fr/Sites-thematiques/Langue-francaise-et-langues-de-France/La-DGLFLF. Accessed October 27, 2020.

Corpus linguistics began to appear before the 1990s. One of its main objectives was the construction of general language dictionaries. Specialised texts were thus excluded as the corpora had to represent the general language. According to Sinclair (2005, 14), "[a]nother factor that affects balance is the degree of specialisation of the text, because a specialised text in a general corpus can give the impression of imbalance". However, the methods developed to study general corpora were subsequently used for specialised corpora: "[c]orpus-based terminology can best be characterized as a working method which explores a collection of domain-specific language material (corpus) to investigate terminological issues" (Gamper and Stock 1998, 149). This is the case for definitional contexts or conceptual relationship markers (Auger and Barrière 2008; L'Homme and Marshman 2006). And indeed, part of corpus linguistics was quickly devoted to the use of specialised corpora to build terminologies. The use of corpora led to a rapprochement between terminologists and linguists. For terminologists, it resulted in the discovery of term variation and the need to propose theories that were more in line with these variations. Sager (1990, 58) writes: "[t]he increasing tendency to analyse terminology in its communicative, i.e., linguistic context, leads to a number of new theoretical assumptions and also new methods of compilation and representation".

One of the main aspects investigated concerned the social dimension of terminology (see chapter by Gaudin and Delavigne in this volume). The study of specialised texts means that it is necessary to adopt a situated vision of how language functions, in other words, to consider text production and its context. For this reason, terminology could no longer be based on a purely onomasiological approach which mainly focused on the creation of norms. Researchers throughout the world began to criticize Wüster's vision and decided to study real term usage, (e.g., Condamines 1995; Gaudin 1993, 2003; Temmerman 2000). For example, Cabré (1999, 114) argued that "Terminology, as both a discipline and an activity, but

especially as a discipline, needs a new orientation which stresses its social and pragmatic aspects."

In Textual Terminology, as for the French semanticist Rastier (1998), it is neither the term nor the corpus that constitutes the most significant element but rather the text (see Chapter by Pecman in this volume). It is mostly at this level that the communicative situation, the field and especially genre should be considered. If texts are collected to form a study corpus, it is because they belong to genres that are relevant to the purpose of the study. In such an approach, the interpretation of language data needs to take the nature of the texts and the objective of the study, into consideration.

In addition to the crucial role of texts, the other defining characteristic of Textual Terminology is the use of NLP and Knowledge Engineering to analyse corpora and structure terminologies.

2.3. Terminology, NLP and Knowledge Engineering

The 1980s witnessed the approximation of terminology and knowledge (or Artificial Intelligence). In 1985, Nedobity wrote a paper entitled "Terminology and Artificial Intelligence". In 1987, the first Terminology and Knowledge Engineering (TKE) conference took place at the University of Trier. In 1991, Skuce and Meyer wrote an article entitled "Terminology and Knowledge Engineering: Exploring a Symbiotic Relationship" (1991, 29/1). In 1993, the Terminology and Artificial Intelligence group in France brought together computer scientists, linguists and terminologists, who played a major role in the development of Textual Terminology. These few examples show that, as early as the end of the 1980s, the close relationship between the issues of terminology and those of Knowledge Engineering (or Artificial Intelligence) was perceived.

Clearly, these disciplines were all united by a shared interest in representing knowledge in the form of networks of terms/concepts. Knowledge engineers realized that they could use existing terminology databases and/or methods inspired by corpus-based terminology (Aussenac-Gilles, Bourigault and Condamines 1995; Aussenac-Gilles, Biélow and Szulman 2000; Buitelaar and Cimiano 2008), and terminologists realized that they could use the knowledge representation modes of knowledge engineering to help build coherent and comprehensive resources (Meyer et al. 1992). In fact, the representation of knowledge in the form of a graph was perfectly adapted to natural language processing tools.

With the development of ontologies, the relationship with terminology became even clearer, which led to the emergence of termino-ontological resources (Bourigault, Aussenac-Gilles and Charlet 2004) or termontology resources (Kerremans, Temmerman and De Baer 2008). In regard to terminology, the evolution of databases enriched with insights from Knowledge Engineering tools gave rise to terminology knowledge bases. The main objective of this representation model is to replace, insofar as possible, a definition by a network of conceptual relationships (Condamines 2018).

The need to systematically build a network of concepts from corpora led to the development of a significant number of studies and tools based on the notion of markers of conceptual relations (and knowledge-rich contexts (Meyer 2001; Marshman in this volume)). Based on this same objective, NLP and Terminology pursued their own research agendas. For NLP, it was a question of data extraction and retrieval though focused on the identification of conceptual relations (Bourigault, L'Homme and Jacquemin 2001). For terminology, it was a question of seeing how to adapt lexicological studies on the construction of local grammars with a view to identifying conceptual relations in corpora. In relation to the phenomena observed in specialised corpora, various studies focused on the role of textual genre and the effectiveness of markers (Condamines 2002, 2008; Marshman, L'Homme and Surtees 2008).

3. The Principles of Textual Terminology Today

This section describes the methodological principles underlying the approach of Textual Terminology, the stages involved as well as the main protagonists.

3.1. Steps in a Textual Terminology analysis

In Textual Terminology, linguistic analysis typically takes place in six stages: (i) definition of the (theoretical or applied) purpose of the analysis; (ii) compilation of the corpus; (iii) selection of tools; (iv) use of tools; (v) validation of the results; (vi) integration of the results into linguistic paradigms.

3.1.1. Purpose of the analysis

The purpose of the analysis may stem from an extra-academic need or from scientific inquiry, often triggered by an extra-academic demand. In any case, especially when it concerns an extra-academic demand, the request needs to be 'translated' into a linguistic question. This makes it possible, on the one hand, to use knowledge already acquired in linguistics and, on the other hand, to integrate the results acquired into the linguistic paradigm of choice once the analysis has been completed. It is then a question of understanding how studies that are doubly situated (by the objective and by the nature of the corpus studied) contribute to the advancement of knowledge.

For example, a few years ago, to celebrate the 50th anniversary of its creation, the CNES (Centre National d'Etudes Spatiales, the French National Space Agency) asked us to carry out a research project on how space, as a body of knowledge, real or imaginary, was more or less consciously disseminated to the general public. We proposed to address this question from the perspective of determinologisation, i.e., the movement of terms from

specialised language into general language (Meyer and Mackintosh 2000). This seemed a good indicator of the penetration of the domain-specific knowledge into general knowledge. A thesis is in progress on this question – practically unexplored in linguistics to date – in the field of particle physics (Humbert-Droz in progress).

3.1.2. Compilation of the corpus

At the beginning of corpus linguistics, a corpus was intended to represent a (sample of) language, mainly with a view to building a general language dictionary or even a grammar. To ensure their representativeness, well-balanced corpora were built to report on a variety of communicative situations, but not specialised communication (see Section 1.2). Even though representativeness might seem easier to achieve in a specialised corpus, the notion of 'domain', does not really simplify the problem of corpus compilation. This notion is generally based on a top-down view. In other words, knowledge is organized *a priori* into domains within which discourse and terminology are assumed to be structured. This division is already problematic, as knowledge is not naturally discontinuous (Rogers 2013).

Moreover, this pre-established organisation is only partially relevant when it comes to considering a particular need. From the point of view of Textual Terminology, the corpus should be as representative as possible, but this representativeness also encompasses the objective of the study (Humbert-Droz, Picton and Condamines 2019). In the case of the previously mentioned determinologisation study, the question of representativeness also takes into consideration the types of texts through which terms are disseminated. We thus chose to build a corpus composed of four sub-corpora: (i) a corpus presenting projects carried out at CNES; (ii) a corpus of press releases issued by the CNES; (iii) a corpus of press releases containing space terms; (iv) and a corpus from the general press (Condamines and Picton 2012).

Another important point within Textual Terminology concerns the comparison of sub-corpora. Depending on the study to be conducted, the corpus is organised into sub-corpora, which are compared in a way that is relevant to the study. Corpora can thus be organised based on time period, levels of expertise, disciplines, different departments within a company, etc.

One difficulty with building a specialised corpus concerns the availability of the texts and the difficulty of accessing data. There are various reasons for this, some of which are specific to the analysis of language for specific purposes. For instance, as Ghaddar and Langlais (2020, 3600) note, in an industrial context, "privacy issues and the commercial value of financial domain data" make it difficult to access data in specialised fields. Grabar et al. (2019) also discuss individual privacy and data protection issues that limit access to certain kinds of medical texts. Moreover, in order to systematically study the productions with the help of tools, texts must be in written and digital form. This requirement excludes, for example, handwritten notes, radio programs (unless they are transcribed), lectures, or even confidential publications. These constraints can result in rather small corpora and create methodological problems (see for instance Bowker and Pearson 2002).

3.1.3. Selection of tools

Tools analyse character strings and not words as semantic units. They quantify these strings, find similarities between them, and/or search for predefined strings in a corpus. The objective of NLP researchers is to propose results that could be associated with a certain meaning. In Textual Terminology, the tools are designed for a specific task (i.e., terminology and conceptual relation extractors). They may be more generic (e.g., taggers, concordancers, lexicometric, textometric tools, etc.). It is therefore necessary to decide which tools to choose, depending on the objective of the study.

3.1.4. Use of tools

Since Textual Terminology proposes to enter the corpus *via* terms, for whatever purpose, a term extractor should be the first tool used. Once the initial list of candidate terms has been established, the next step is to analyse the contexts in which they appear. There are two ways to study such contexts. The first way is to study the meaning based on the distribution of terms in the corpus (bottom-up analysis). The second way is to search term contexts for elements that are known to be relevant to the study (top-down analysis).

Bottom-up analysis; This type of analysis involves categorizing term contexts and then interpreting them (based on our competence as speakers and linguists). It is based on the well-known distributional approach developed in two currents, one inspired by sociolinguistics (Firth 1957) and the other by a mathematical (and behaviourist) vision of language (Harris 1951). The main idea underlying this approach is that there is a correlation between distributional similarity and meaning similarity. It is thus possible to understand the meaning of a word or group of words by considering all its contexts.

Most analyses in Textual Terminology focus on corpora and the comparison of sub-corpora. In such cases, differences in distribution are associated with differences in the communicative situation, prior to the sub-corpus organisation. For example, if the corpus is organised in sub-corpora of experts vs. non-experts (e.g., to study popularisation as in Delavigne [2017] or Ledouble [2020]), a difference in term distribution, such as when the term is used in general language, can be interpreted as a sign of meaning variation.

Top down analysis; In the early days of Textual Terminology, when the main purpose was the construction of terminology networks, research focused on defining and specifying conceptual relation markers or lexico-syntactic elements to which an *a priori* interpretation is associated. The hypothesis is that these markers, associated with *knowledge-rich contexts*

(Meyer 2001), are less likely to vary from one specialised corpus to another. The choice of appellation is very interesting as it introduces the notion of "wealth of knowledge". This is not only relevant to conceptual relation markers but also to linguistic units of different types, related to different types of knowledge. This means that for each analysis, elements that are directly related to the objective of the study can be defined *a priori*. The purpose here then is not to build a relational network, but rather to track the presence of elements that indicate a shift in meaning, variation, etc. that speakers are aware of.

For example, if the analysis focuses on the evolution of a term, word, or phrase over a period of time, one can find markers such as, *before*, *[concept name2] formerly called [concept name1]* (Picton 2009, 2014), that "mark" the presence of relevant information regarding the evolution of a domain or its terminology.

3.1.5. Validation of the results

In Textual Terminology, the results obtained must be validated at many levels. Firstly, results must be acceptable to peer analysts. This implies that the proposed interpretations must be associated with textual contexts that justify them. These interpretative choices can be discussed and justified if necessary.

The second type of validation is the one provided by experts in the field. One of the characteristics of specialised corpus analysis is that the analyst in most cases is not a domain expert and is thus not familiar with the terms or phrasing used by experts. This means that data may be misinterpreted. It is therefore necessary to work closely with experts in the field, who analyse a set of contexts that have been pre-selected either because of their difficulty or importance. Together, they can thus co-construct an interpretation that makes use of linguistic knowledge and knowledge of the field. Validation by experts then often proceeds in a co-construction process (Picton 2009, 2014).

A third type of validation can be performed by end-users, when the project call stems from an applied need. From a strictly lexical point of view, these users are language professionals, such as translators, technical writers, or even computer scientists who need resources to make the tools work. Nevertheless, end-users can also be LSP teachers, non-experts wishing to access simplified definitions, experts from other disciplines, etc.

Therefore, the results are not only a set of terminology entries. They can also be data interpretations based on the study of terms. The way in which these results are presented may fall under what, in ergonomics, is called ‘usability’. This part of the validation can therefore be performed in association with ergonomists, which also situates Textual Terminology within the framework of ergonomic linguistics (see Section 4.1.2.).

3.1.6. Integrating the results into linguistic paradigms

From a theoretical point of view, the consideration of extra-academic needs presents a risk. Even though the results might fulfil these needs in an *ad hoc* manner, their contribution to linguistics might be overlooked. On the one hand, the results obtained might not be sufficient to enrich the description of terms and their use. In fact, it goes without saying that if results are not elaborated, they cannot be reused to deal with other similar problems. For instance, in the study of determinologisation, the results obtained are relevant to neology. In the case of popularisation, they pertain to shifts in meaning (e.g., Condamines, Humbert-Droz and Picton, forthcoming or Ledouble 2020). Integrating (and theorising) these results in linguistic paradigms is thus of major importance.

3.2. Skills and protagonists

Echoing what has been described above, the following types of protagonist may be involved in this process (Table 1):

- Terminologists-linguists, who organize the work and conduct the analyses. From the perspective of a scientific study, they act as intermediaries between the extra-academic world and linguistic knowledge.
- Domain experts, who help to compile the corpus, validate the interpretation of the results and, in some cases, validate the results within the context of the initial project goals.
- End users, who may work in companies and are often the source of the demand, driven by quality departments or design offices.
- Terminologists-linguists may call upon other types of skill, e.g., those of NLP experts to adapt computer tools to the needs of the analysis.
- Ergonomists and their expertise can be used to characterize the need and evaluate the usability of the proposed results.

Table 1 – Tasks/stages and protagonists - Textual Terminology

Tasks/Stages	**Protagonists**
Identification of a need or a problem (translation into a linguistic issue)	(ergonomists) End-users Domain experts Language experts (analysts, terminologists-linguists)
Constitution of the corpus	Language experts Domain experts
Selection and use of tools	Language experts If necessary, NLP experts
Interpretation of the clues provided by tools	Language experts
Construction of a final interpretation	Language experts Domain experts
Integration of results into a linguistic	Language experts

paradigm	
Validation	(Ergonomists) Domain experts End-users

Since the early 1990s, these principles have remained relatively stable. However, several aspects need to be re-examined. On the one hand, a major evolution has taken place in NLP and Knowledge Engineering, thanks to the availability of huge amounts of data on the Internet. On the other, studies of specialised corpora have diversified. The following section describes the evolution of Textual Terminology and the perspectives that have emerged.

4. Textual Terminology today: Challenges and Perspectives

Over the last thirty years, Textual Terminology has evolved considerably in regard to the following: (i) emergence of applied needs; (ii) development of corpus linguistics; (iii) cross-fertilisation with NLP, KE/AI. This section reappraises these three very closely linked areas, as they are today. This sheds light on the future of Textual Terminology and its development in the short and medium term.

4.1. New needs and perspectives

4.1.1. New needs in Terminology

Whereas research initially focused on the creation of termino-ontological resources (see Section 1), from the early 2000s onwards, the need to update these resources began to surface (e.g., Condamines, Rebeyrolle and Soubeille 2004). This question led to the consideration of diachronic variation (see Section 2).

From then, new needs gradually emerged, which involved the study of other types of dialogical variation (Freixa 2006). As explained in section 2.1.1, a project request from the

CNES led to the integration of the analysis of diastratic variation. At the same time, another project was initiated by the Geneva Centre for Humanitarian Studies[2]. Its objective was to highlight differences in term usage from different perspectives in the field, in English[3] (Egger, Picton and Schopper 2018). Diatopic variation is in this case important. We can also mention a study that we carried out on exobiology (funded by the CNES), an emerging discipline that involves different disciplines. This work shed light on the fact that terminological variation was not only a difficulty for interdisciplinarity, but that it could contribute to identifying the dynamics between multiple points of view regarding a concept (Condamines 2014).

In addition to envisaging other types of variation, this led us to reflect on the simultaneous analysis of different kinds of variations. In the case of determinologisation, for example, variation is both the diastratic (movement of terms from specialised texts to general texts) and the diachronic (movement of terms over time). This implies working on fairly complex comparable corpora, which are analysed, based on different types of variation.

In these analyses, the general methodological approach remains the same as the one described in Section 3. However, in addition to the overall difficulty of ensuring the representativeness of the data and their balance in the different sub-corpora (Section 3.1.1.), the consideration of these more complex comparable corpora also highlights the issue of data availability. This question is directly related to the NLP tools and methods that Textual Terminology can use with these corpora (see Section 4.2.).

4.1.2. Towards new kinds of need: Controlled Natural Languages

As a method of systematic linguistic analysis with NLP tools in specialised domains, Textual Terminology can be implemented in different ways. Its methodology can also be adapted to

[2] https://humanitarianstudies.ch/. Accessed October 17, 2020. (the Centre was formerly known as *CERAH*).
[3] https://evidenceaid.org/do-we-all-speak-and-practice-humanitarian-in-the-same-way/. Accessed October 17, 2020.

elements other than lexical units, such as the definition of Controlled Natural Languages (CNLs). CNLs are linguistic recommendations intended to limit risks related to the use of the natural language, particularly in companies and businesses (Ryan 2018, 294). Since this is also one of the functions of terminologies (Condamines 2010), many CNLs also have a terminology component. Like terminologies before the 90s, they are usually developed by experts in the field, sometimes with the help of translators. This means that CNLs are mainly established by introspection, based on the competence of domain experts.

In another project with the CNES, we were asked to build a CNL for the writing of satellite specifications. We thus compiled a corpus of specifications written without instructions in order to identify their lexico-syntactic regularities (typical of that textual genre) and to use these regularities to construct the CNL.

We used the SDMC tool (Sequential Data Mining under Constraints) (Quiniou, Cellier and Legallois 2012, 167), on the corpus of specifications from the CNES. The tool was able to identify phrasemes, which not only include recurring lexico-syntactic structures but also multiword terms (for example, *en pointage inertiel* [inertial pointing]) and expressions. By deleting the terms present in the CNES database, we were able to maintain the structures that seemed to correspond to sentence structures (Warnier and Condamines 2015), for example [*sur réception de tc, le lvc [envoyer]* ...] (on receipt of the tc, the lvc [send]...). Based on the selected structures, we proposed alternative structures that conveyed the same information. In the example above, *envoyer* [send] may have the following forms: *envoie/doit envoyer/doit pouvoir envoyer/enverra* [send/must send/must be able to send/ will send]. Concerning injunctions, 20 sets of structures were thus specified in different lexical contexts. These structures were then proposed to specialists in the writing of specifications, who evaluated their acceptability. These evaluations of spontaneous formulations (corpus) by means of

acceptability tests (Warnier 2018) allowed us to propose usable recommendations. In fact, CNES decided to use our CNL to train the writers and editors of specifications.

These results point to the need for companies and businesses to employ linguists. The development of this CNL also led to a number of theses carried out at the CLLE laboratory and highlighted the need to borrow concepts such as ‘usability’ from ergonomics. The consideration of usability in the construction or adaptation of CNLs allowed Condamines (2020) to develop the concept of ‘ergonomic linguistics’.

Ergonomic linguistics provides a framework to propose answers to language needs, especially within firms. The aim is to propose linguistic standards (terminologies, CNLs, etc.) that will be effective in a given context as well as acceptable to users. From the perspective of Textual Terminology, (see Table 1), at the beginning of the study, ergonomic linguistics characterizes a need, and at the end of the analysis, it evaluates the relevance of the results in relation to this need and helps to make them available in a form that can be applied by the end-users. The integration of ergonomic skills and perhaps even cognitive psychology (to conduct usability testing) could be an interesting perspective for Textual Terminology.

4.2. Textual Terminology, NLP, and KE/AI: Where does the “symbiotic relationship” go?

As previously mentioned in Section 1, corpus analysis for the compilation of termino-ontological resources was one of the founding pillars of Textual Terminology and of its connection with NLP and KE. However, the situation is quite different today. According to Condamines (2018, 343), "the original 'symbiotic relationship' between terminology and Knowledge Engineering, evoked by Skuce and Meyer (1991), no longer applies". There are at least two possible reasons for this: (i) the loss of a common research objective; (ii) the question of corpus size and the methods used to explore it.

4.2.1. From symbolic approaches to machine learning: data quantity

As pointed out by Condamines (2018), research objectives have experienced a dramatic change. There is considerably less emphasis placed on building termino-ontological resources from texts, especially based on symbolic approaches, whereas more is placed on ontologies, based on automatic learning methods. The development of these methods has triggered a real paradigm shift in NLP. Even if some studies still use a symbolic approach to build ontologies (e.g., Asim et al. 2018), studies such as those by Church (2011) or Hall, Jurafsky and Manning (2008) have estimated that since the early 1990s, the use of statistical methods in NLP (including machine learning) has risen from 30% to over 90%.

The role of linguistics in NLP has thus become less important. As Eensoo and Valette (2013, 2), point out, "[p]our une même application – par exemple la traduction automatique – les méthodes symboliques, jadis, mobilisaient pendant plusieurs années une armada de linguistes pour l'écriture de règles ; aujourd'hui, pour peu que des corpus parallèles de taille suffisante soient disponibles, un système par apprentissage nécessitera très peu de ressources humaines et de temps".[4] Efforts in NLP are now focused on developing these approaches, which most often require very large corpora.

Despite the lack of consensus on corpus size (e.g., Corpas Pastor and Seghiri Domínguez 2010), Fabre and Lenci (2015, 9) observe, "[t]here has been a clear shift from the treatment of middle-sized specialised corpora to the acquisition of distributional thesauri in the 90's (…), to the compilation of corpora as large as possible, often heterogeneous in genre and domain". In the context of (neural) machine translation for instance, Ghaddar and Langlais (2020, 3595) explain that "[t]hese models require the use of large-scale parallel

[4] "[f]or a single application - for example automatic translation - symbolic methods used to mobilize an armada of linguists for several years to write rules; today, provided that parallel corpora of sufficient size are available, a learning-based system will require very few human resources and time" (our translation).

corpora to train millions of internal parameters". In the field of automatic distributional analysis methods, underlying much current research in NLP (e.g., Heylen and Bertels 2016; Lenci 2018), Boleda (2020, 230) specifies that "[a] rule of thumb is to have at least 20-50 instances of each expression to represent; many languages, domains, or time periods simply lack these data."

Whereas some authors seek to show that applications of current distributional approaches are possible in smaller corpora[5] (e.g., Fabre et al. 2014; Tanguy, Brunet and Ferret 2019), these corpora still have a minimum of several million occurrences, which is still quite large in comparison to the usual size of specialised corpora (e.g., L'Homme 2020).

Today some (non-exhaustive) examples of large specialised corpora include the following: more than 23M occurrences for the Ecolexicon corpus (León-Araúz, San Martín and Reimerink 2018); 85M for the Humanitarian Encyclopedia[6] (Egger, Picton and Schopper 2018); 4.5M for the TALN corpus built by CLLE lab (Tanguy, Fabre and Bard 2020), etc. Moreover, with the development of new demands for terminology analysis (Section 3.1.), an increase in the size of specialised corpora is foreseeable. However, the following points about large specialised corpora should be highlighted:

- Specialized corpora are still relatively small in comparison to current approaches in machine learning (and especially distributional analysis in general language).
- Even if the total number of occurrences is in the billions, the division into sub-corpora for the purposes of comparability considerably reduces the size of the datasets to be analysed. In the case of the Humanitarian corpus (85M occurrences), the sub-corpora have between several millions and less than 1M occurrences (Picton, Drouin and Humbert-Droz, forthcoming).

[5] Special issues of periodic journals also question the place and assets of small corpora, for instance the CORPUS journal, issue 18, https://journals.openedition.org/corpus/3094. Accessed October 17, 2020.
[6] https://humanitarianencyclopedia.org/concept-analysis/. Accessed October 17, 2020.

- Specialized corpora are still difficult to collect and compile (section 2.1.1).

Nevertheless, in view of new needs and the fact that data access is, to a certain extent, easier, Textual Terminology should tackle the issue of large specialised corpora, especially since it is closely linked to that of the NLP methods and approaches.

4.2.2. Analysing specialised corpora today: methodology

The nature and evolution of specialised corpora raise different questions regarding the tools and NLP methods available for the analyses, and in particular, as they require the combination of quantitative analysis (e.g., to use data to identify leads and trends) and qualitative analysis (to refine the initial quantitative observations and interpret them appropriately).

This issue goes hand in hand with the need for data visualisation to analyse complex data (Picton, Drouin and Humbert-Droz, forthcoming). Caple, Bednarek and Anthony (2018), for example, discuss this issue in relation to the analysis of data from different media (including text and images in news items). Anthony (2018) shows the diversity of possible means of visualisations and their impact on corpus analysis results. In diachronic linguistics, Hilpert (2011) proposes dynamic visualisation solutions to handle big diachronic comparable corpora for a fine-grained description of phenomena. Although these reflections seem to largely apply to the field of general language, Textual Terminology can greatly benefit from them.

Nonetheless, other approaches in NLP research seem far removed from the specific needs of Textual Terminology and specialised corpora. For example, distributional analysis approaches and techniques, which are at the heart of Textual Terminology methods (Section 3.1.4.), are developing exponentially in NLP, along with machine learning and big data possibilities. In terminology, this path has been explored in some depth for semantic relation extraction (e.g., Bernier-Colborne and Drouin 2016, Bertels, this volume) and term extraction

(e.g., Drouin, Morel and L'Homme 2020). However, regarding the overall corpus analysis, two issues arise:

- The complexity and diversity of distributional approaches make it often difficult for terminologists to understand how these approaches work and their applicability to the data at their disposal (e.g., Boleda 2020). Distinguishing the pros and cons of each approach requires advanced skills related to computer science, AI, and statistics.
- As previously mentioned, these methods are generally based on very large and heterogeneous general data, which do not reflect the context and needs of specialised corpora, even large specialised corpora. Current discussions in NLP do not therefore address certain issues that are directly related in terminology.

However, various studies tend to show that the nature of corpora has an impact on the results provided by a machine learning approach in distributional analysis. For example, Del Tredici , Fernández and Boleda (2019) discusses recent methods for distributional analyses and short-term diachrony, whose relevance for terminology is evident (e.g., Dury 2004, in this volume; Picton 2014). These authors clearly highlight the impact of the temporal shift in corpora on the results of diachronic distributional analysis and the difficulties of interpreting them. Choosing the right distributional method for the right corpus is thus of prime importance but remains a difficult point that requires navigating between the specific needs of terminology and the evolution of NLP.

Regarding this distance between NLP and Textual Terminology, Firth's (1957) remark, "[y]ou shall know a word by the company it keeps", is very often cited to justify the distributional approach in NLP. This quote is often taken out of context, however, with respect to Firth's work as a whole. In another excerpt, quoted in Duvivier-Senis (2016), Firth describes the distributional approach as follows: "[f]irst the structure of **the appropriate**

contexts of situation must be stated[7]. Then the syntactical structure of the texts. The criteria of distribution and collocation should then be applied" (Firth [1952] 1968, 19). Various authors emphasise the same idea concerning the interpretation of the results provided by distributional approaches, such as in Fabre and Lenci (2015, 9):

> What was clearly asserted in Harris' original method was the fact that such inductive semantic classifications reflected the Distributional Semantics Today's use of words in specific corpora. The approach was set in the context of the theory of sublanguages, based on the assumption that only **corpora from restricted domains** could guarantee the possibility to build up clear-cut semantic categories. (our emphasis)

Consequently, there seems to be a certain distance between the original assertions by Firth and Harris (who had, however, different theoretical perspectives) on distributional analysis and current developments in machine learning and Artificial Intelligence on big data, which rely on large and heterogeneous data, by definition.

These elements fuel the reflection about the nature of the relationship between NLP and Textual Terminology today. While it is clear that the symbiotic relationship no longer exists, close collaboration seems possible and even essential. However, this collaboration must consider the specificities of specialised (large and complex) corpora and question the relevance and portability of technical developments in NLP for the new needs of Textual Terminology.

5. Concluding remarks

This chapter has proposed a retrospective look at the origins of Textual Terminology. After describing its methodological approach and the questions it raises, this perspective enabled us to highlight the evolution of this theory in the short and medium term, and where to situate it today.

[7] Our emphasis.

The elements used to define Textual Terminology might initially seem quite similar to those of applied corpus linguistics (Hyland, Chau and Handford 2012) or even specialised discourse analysis (Gotti and Giannoni 2006). However, the specificity of Textual Terminology lies in the following: (i) it concerns only specialised corpora; (ii) the purpose of the study is considered throughout the analysis; (iii) terms are the point of entry into the texts; and (iv) tools are used as much as possible, with particular attention to the nature of the results obtained.

It also allowed us to highlight the cross-fertilisation between Textual Terminology and other theories of terminology and, consequently, the ideas they share. We are referring to the need to base analysis on specialised corpora, the consideration of the situational context of project calls, the consideration of polysemy and variation, etc. (see chapter by Delavigne and Gaudin in this volume). Nevertheless, Textual Terminology, while nourishing these links, bases its specificity on four main axes: (i) the focus on tools for exploring corpora; (ii) the systematisation of its methodological approach (with a view to scientificity and reproducibility); (iii) the central role of experts (language professionals, users, field experts, ergonomists) throughout the work process; (iv) applicability of results to linguistics.

This picture of Textual Terminology today, as well as of its needs and challenges, led us to highlight the new ways in which it can work together with other disciplines. For instance, within the framework of Textual Terminology, ergonomics and psychology have become of great importance to meet new needs (e.g., CNLs) and ensure the usability of the results obtained. Finally, the relationship of Textual Terminology with NLP and KE/AI has been transformed by the scientific goals and developments specific to each discipline. From being symbiotic, the modalities of this relationship have evolved into a real need for collaboration. Because of these expanded horizons, Textual Terminology is currently in rapid development and can look forward to a very bright future.

References

Asim, Muhammad Nabeel, Muhammad Wasim, Muhammad U.G. Khan, Waqar Mahmood, and Hafiza Mahnoor Abbasi. 2018. "A Survey of Ontology Learning Techniques and Applications." *Database* Vol. 2018, bay101. https://doi.org/10.1093/database/bay101.

Anthony, Lawrence. 2018. "Visualisation in Corpus-based Discourse Studies." In *Corpus Approaches to Discourse: A Critical Review*, edited by Taylor, Charlotte and Anna Marchi, 197-224. London: Routledge.

Auger, Alain, and Caroline Barrière (eds.). 2008. "Pattern-based Approaches to Semantic Relation Extraction: A State-of-the-art." *Terminology* 14 (1): 1-19.

Aussenac-Gilles Nathalie, Didier Bourigault, and Anne Condamines. 1995. "How can Knowledge Acquisition Benefit from Terminology?" In *Proceedings of the 9th Knowledge Acquisition Workshop, Advances in Knowledge Acquisition*, Banff, Alberta, edited by Brian R. Gaines and Mark Musen, 1-19.

Aussenac-Gilles, Nathalie, Brigitte Biébow, and Sylvie Szulman. 2000. "Revisiting Ontology Design: A Method Based on Corpus Analysis." In *Proceedings of the European Knowledge Acquisition Workshop, EKAW 2000: Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, Juan-les-Pins, France, edited by Rose Dieng and Olivier Corby, 172–188.

Bernier-Colborne, Gabriel, and Patrick Drouin. 2016. "Evaluation of Distributional Semantic Models: a Holistic Approach." In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, Osaka, Japan, edited by Patrick Drouin, Natalia Grabar, Thierry Hamon, Kyo Kageura and Koichi Takeuchi, 52–61. COLING.

Boleda, Gemma. 2020. "Distributional Semantics and Linguistic Theory." *Annual Review of Linguistics* 6 (1): 213–234. doi:10.1146/annurev-linguistics-011619-030303.

Bourigault, Didier, Nathalie Aussenac-Gilles, and Jean Charlet. 2004. “Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas.” *Revue d'Intelligence Artificielle (RIA)* 18: 87–110.

Bourigault, Didier, Marie-Claude L'Homme, and Christian Jacquemin (eds.). 2001. *Recent Advances in Computational Terminology*. Amsterdam/ Philadlephia: John Benjamins.

Buitelaar, Paul, and Philip Cimiano. 2008. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge.* Amsterdam: IOS Press.

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

Cabré, Maria Teresa. 1999. *Terminology Theory, methods and applications*. Amsterdam/ Philadelphia: John Benjamins.

Caple, Helen, Monika Bednarek, and Lawrence Anthony. 2018. “Using Kaleidographic to Visualize Multimodal Relations ithin and across Texts.” *Visual Communication* 17 (4): 461-474. doi:10.1177/1470357218789287.

Church, Ken W. 2011. “A Pendulum Swung too Far.” *Linguistic Issues in Language Technology* 6: 1-27.

Condamines, Anne. 1995. “Terminology: New Needs, New Perspectives.” *Terminology 2*(2): 219-238.

Condamines, Anne. 2002. “Corpus Analysis and Conceptual Relation Patterns.” *Terminology* 8 (1): 141-162.

Condamines, Anne. 2008. “Taking *Genre* into Account for Analyzing Conceptual Relation Patterns.” *Corpora* 8: 115-140.

Condamines, Anne. 2010. “Variations in Terminology. Application to the Management of Risks Related to Language Use in the Workplace.” *Terminology* 16 (1): 30-50.

Condamines, Anne. 2014. “How Can Linguistics Help to Structure a Multidisciplinary Neo-Domain Such as Exobiology?” *BIO Web of Conferences* 2, 06001. https://www.bio-conferences.org/articles/bioconf/abs/2014/01/bioconf_epov2012_06001/bioconf_epov2012_06001.html.

Condamines, Anne. 2018. “Terminological Knowledge Bases.” In *The Routledge Handbook of Lexicography*, edited by Pedro Fuentes, 335-349. London: Routledge.

Condamines, Anne. 2020. “Towards an Ergonomic Linguistics. Application to the Design of Controlled Natural Languages.” *International Journal of Applied Linguistics*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1111/ijal.12313.

Condamines, Anne, Julie Humbert-Droz, and Aurélie Picton. Forthcoming. “Néologie par déterminologisation : méthode de repérage et catégorisation en corpus dans le domaine de la physique des particules.” In *La néologie des langues romanes : nouvelles dynamiques et enjeux*, edited by Maria Bélen Villar Díaz, José Carlos de Hoyos, Pascaline Dury, Julie Makri-Morel, and Vincent Renner. Bern: Peter Lang.

Condamines, Anne, and Aurélie Picton. 2012. “Des communiqués de presse du Cnes à la presse généraliste. Vers un observatoire de la diffusion des termes.” In *Langues de spécialité et néologie*, edited by Pascaline Dury, José Carlos de Hoyos, Julie Makri-Morel, François Maniez, Vincent Renner and Maria Belén Villar Díaz, 141-161. Lyon: CRTT.

Condamines, Anne, Josette Rebeyrolle, and Anne-Marie Soubeille. 2004. “Variation de la terminologie dans le temps : une méthode linguistique pour mesurer l'évolution de la connaissance en corpus.” In Proceedings of *EURALEX 2004*, Lorient, France, edited by Geoffrey Williams and Sandra Vessier, 547-557. Lorient: Université de Bretagne Sud.

Corpas Pastor, Gloria, and Miriam Seghiri Domínguez. 2010. "Size Matters: A Quantitative Approach to Corpus Representativeness." In *Lengua, traducción, recepción: en honor de Julio César Santoyo*, edited by Rosa Rabadán, Marisa Fernández López and Trinidad Guzmán González, 111-145. Vol. 1. Léon: Publicaciones Universidad de León.

Del Tredici, Marco, Raquel Fernández, and Gemma Boleda. 2019. "Short-Term Meaning Shift: A Distributional Exploration". In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, edited by Jill Burstein, Christy Doran and Thamar Solorio, 2069-2075. Stroudsburg: Association for Computational Linguistics.

Delavigne, Valérie. 2017. "Term Usage and Socioterminological Variation. The Impact of Social and Local Issues on the Movement of Terms." In *Multiple Perspectives in Terminological Variation*, edited by Patrick Drouin, Aline Francoeur, John Humbley and Aurélie Picton, 31-56. Amsterdam/ Philadelphia: John Benjamins.

Drouin, Patrick, Jean-Benoît Morel, and Marie-Claude L'Homme. 2020. "Automatic Term Extraction from Newspaper Corpora: Making the Most of Specificity and Common Features." In *Proceedings of 6th International Workshop on Computational Terminology*, Marseille, France, edited by Béatrice Daille, Kyo Kageura and Ayla Rigouts Terryn, 1-7. Paris: ELRA.

Dury, Pascaline. 2004. "Building a Bilingual Diachronic Corpus of Ecology: The Long Road to Completion." *ICAME Journal* 28: 5-16.

Duvivier-Senis, Angela. 2016. "John Rupert Firth, Historien de la linguistique et fondateur de la 'London School'." PhD diss., Université de Paris 7. http://theses.md.univ-paris-diderot.fr/SENIS_Angela_2_va_20161118.pdf.

Eensoo, Egle, and Mathieu Valette. 2013. *Sémantique textuelle et TAL : un exemple d'application à l'analyse des sentiments*. Working paper. https://halshs.archives-ouvertes.fr/halshs-00968634.

Egger, Clara, Aurélie Picton, and Doris Schopper. 2018. "How Do Humanitarian IO Conceptualise the World ? An Interdisciplinary Approach to the Analysis of Concepts Used in Humanitarian Action." Paper presented at the RUN Workshop, Geneva, Switzerland.

Fabre, Cécile, Nabil Hathout, Franck Sajous, and Ludovic Tanguy. 2014. "Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille." In *Actes de la 21^{e} conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014) Atelier Sémantique Distributionnelle (SemDis 2014)*, Marseille, France, June 2014, edited by Cécile Fabre, Nabil Hathout, Lydia-Mai Ho-Dac, François Morlane-Hondère, Philippe Muller, Franck Sajous, Ludovic Tanguy, and Tim van de Cruys, 266–279.

Fabre, Cécile, and Alessandro Lenci. 2015. "Distributional Semantics Today: Introduction to the special issue." *Traitement Automatique des Langues* 56 (2): 7-20.

Firth, John Rupert. (1952) 1968. "Linguistic Analysis as a Study of Meaning." In *Selected Papers of J.R. Firth*, edited by Frank Robert Palmer, 12-26. London: Longman.

Firth, John Rupert. 1957. *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press.

Freixa, Judit. 2006. "Causes of Denominative Variation in Terminology: A Typology Proposal." *Terminology* 12 (1): 51-77.

Gamper, Johann and Oliviero Stock. 1998. "Corpus-based terminology." *Terminology* 5 (2): 147–160.

Gaudin, François. 1993. *Pour une socioterminologie - Des problèmes sémantiques aux pratiques institutionnelles*. Rouen: Publications de l'Université de Rouen.

Ghaddar, Abbas and Philippe Langlais. 2020. “SEDAR: a Large Scale French-English Financial Domain Parallel Corpus.” In *Proceedings of the 12th Language Resources and Evaluation, LREC 2020*, Marseille, France, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, et al., 3595–3602. Paris: ELRA.

Gotti, Maurizio, and Davide Simone Giannoni (eds.). 2006. *New Trends in Specialised Discourse Analysis*. Bern: Peter Lang.

Grabar, Natalie, Cyril Grouin, Thierry Hamon, and Vincent Claveau. 2019. “Annotated Corpus with Clinical Cases in French.” In *Actes de la 26th conférence sur le Traitement Automatique des Langues Naturelles (TALN’2019)*, Toulouse, France, 1-14. https://hal.archives-ouvertes.fr/hal-02391878.

Hall, David, Daniel Jurafsky, and Christopher Manning. 2008. “Studying the History of Ideas Using Topic Models.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP’08)*, Honolulu, Hawaii, 363-371. https://web.stanford.edu/~jurafsky/hallemnlp08.pdf.

Harris, Zellig S. 1951. *Methods in Structural Linguistics.* Chicago: University of Chicago Press.

Heylen, Kris, and Anne Bertels. 2016. “Sémantique distributionnelle en linguistique de corpus.” *Langages* 201 (1): 51-64. doi:10.3917/lang.201.0051.

Hilpert, Martin 2011. “Dynamic Visualizations of Language Change: Motion Charts on the Basis of Bivariate and Multivariate Data from Diachronic Corpora.” *International Journal of Corpus Linguistics* 16 (4): 435-461.

Humbert-Droz, Julie. In progress. “Circulation des termes entre langues de spécialité et langue générale : proposition d’un cadre théorique et méthodologique d’analyse du

phénomène de la déterminologisation." (provisional title) PhD diss., University of Geneva and University of Toulouse Jean-Jaurès.

Humbert-Droz, Julie, Aurélie Picton, and Anne Condamines. 2019. "How to Build a Corpus for a Tool-based approach to Determinologisation in the Field of Particle Physics." *Research in Corpus Linguistics* (7): 1-17. doi:10.32714/ricl.07.01.

Hyland, Ken, Meng Huat Chau, and Michael Handford. 2012. *Corpus Applications in Applied Linguistics*. London/ New York: Continuum International Publishing Group.

Kerremans, Koen, Rira Temmerman, and Peter De Baer. 2008. "Construing Domain Knowledge via Terminological Understanding." *Linguistica antverpiensia* 7: 177–191.

Ledouble, Hélène. 2020. "Term Circulation and Conceptual Instability in The Mediation of Science: Binary Framing of the Notions of Biological versus Chemical Pesticides." *Discourse & Communication* 14 (5): 466-488.

Lenci, Alessandro. 2018. "Distributional Models of Word Meaning." *Annual Review of Linguistics* 4 (1): 151-171. doi:10.1146/annurev-linguistics-030514-125254.

León-Araúz, Pilar, Antonio San Martín, and Ariane Reimerink. 2018. "The EcoLexicon English Corpus as an open corpus in Sketch Engine." In *Proceedings of the XVIII EURALEX International Congress*. *Lexicography in Global Contexts,* edited by Jaka Čibej, Vojko Gorjanc, Iztok Kosem, and Simon Krek, 893–901. Ljubljana: Ljubljana University Press.

L'Homme, Marie-Claude. 2020. *Lexical Semantics for Terminology. An Introduction*. Amsterdam/ Philadelphia: John Benjamins.

L'Homme, Marie-Claude, and Elizabeth Marshman. 2006. "Terminological Relationships and Corpus-based Methods for Discovering them: An Assessment for Terminographers." In *Lexicography, Terminology, and Translation. Text-based studies in honour of Ingrid Meye*r, edited by Lynne Bowker, 67-80. Ottawa: University of Ottawa Press.

Marshman, Elizabeth, Marie-Claude L'Homme, and Victoria Surtees. 2008. "Portability of Cause-effect Relation Markers across Specialised Domains and Text genres: A Comparative Evaluation." *Corpora* 3 (2): 141-172.

Meyer, Ingrid. 2001. "Extracting Knowledge-Rich Contexts for Terminography : A Conceptual and Methodological Framework." In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Marie-Claude L'Homme, and Christian Jacquemin, 279-302. Amsterdam/ Philadelphia: John Benjamins.

Meyer, Ingrid, Douglas Skuce, Lynne Bowker, and Karen Eck. 1992. "Towards A New Generation of Terminological Resources: An Experiment In Building A Terminological Knowledge Base." In *Proceedings of the Fifteenth Conference on Computational Linguistics (COLING '92) Vol. 3*, edited by Christian Boitet, 956–930. Stroudsburg, PA: Association for Computational Linguistics (ACL).

Meyer, Ingrid, and Krinten Mackintosh. 2000. "When Terms Move into Our Everyday Lives: An Overview of De-terminologization." *Terminology* 6 (1): 111–138.

Nedobity, Wolfgang. 1985. "Terminology and Artificial Intelligence." *Knowledge Organization* 12 (1): 17-19.

Picton, Aurélie. 2009. "Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial." PhD diss., Université Toulouse Jean-Jaurès.

Picton, Aurélie. 2014. "The Dynamics of Terminology in Short-Term Diachrony: A Proposal for a Corpus-based Methodology to observe Knowledge Evolution." In *The Dynamics of Culture-bound Terminology in Monolingual and Multilingual Communication, Terminology and Lexicography Research and Practice*, edited by RitaTemmerman and Marc Van Campenhoudt, 159–182. Amsterdam/ Philadephia: John Benjamins.

Picton, Aurélie, Patrick Drouin, and Julie Humbert-Droz. Forthcoming. “Exploration et visualisation de la variation terminologique en corpus spécialisés complexes : réflexions et propositions méthodologiques.” In *Lexique(s) et genre(s) textuel(s) : approches sur corpus,* edited by Mathieu Mangeot and Agnès Tutin, 99-116. Paris: Editions des archives contemporaines.

Quiniou, Solène, Peggy, Cellier, Thierry Charnois and Dominique Legallois. 2012. “What about Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics?” In *Computational Linguistics and Intelligent Text Processing*, edited by Alexandra Gelbukh, 166–177. Berlin/ Heidelberg: Springer.

Rastier, François. 1998. “On Signs and Texts: Cognitive Science Faces Interpretation.” *Applied Semiotics/Sémiotique appliquée* 2(5): 195-244.

Rogers, Margaret. 2013. “What is a ‘Domain’ and is this a Useful Question?” *Asp, revue du GERAS* 64: 5-16. doi:10.4000/asp.3810.

Ryan, Richard 2018. “Controlled Language.” In *Languages for Special Purposes: An International Handbook*, edited by John Humbley, Gerhards Budin and Lauren Christer, 289-306. Berlin/ Boston: De Gruyter Mouton.

Sager, Juan Carlos. 1990. *A Practical Course in Terminology Processing*. Amsterdam/ Philadelphia: John Benjamins.

Sinclair, John. 2005. “Corpus and Text - Basic Principles”. In *Developing Linguistic Corpora: A Guide to Good Practice*, edited by Martin Wynne. http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf.

Skuce, Douglas, and Ingrid Meyer. 1991. “Terminology and Knowledge Engineering: Exploring a Symbiotic Relationship.” In *Proceedings of the 6th International Workshop on Knowledge Acquisition for Knowledge-Based Systems*, Banff, Canada, 29/1-29/21.

Tanguy, Ludovic, Pauline Brunet, and Olivier Ferret. 2019. “Extrinsic Evaluation of French Dependency Parsers on a Specialised Corpus: Comparison of Distributional Thesauri.” In *Proceedings of the 12th Language Resources and Evaluation, LREC 2020*, Marseille, France, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, et al., 5820-5828. Paris: ELRA.

Tanguy, Ludovic, Cécile Fabre, and Yoann Bard. 2020. “Impact of Document Structure on Distributional Semantics Models: a Case Study on NLP Research Articles.” In *Actes de la 27e conf/rence sur le Traitement Automatique des Langues Naturelles (TALN’2020)*, Nancy, France, edited by Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, and Stéphane Schneider, 122-135. ATALA and AFCP. https://www.aclweb.org/anthology/2020.jeptalnrecital-taln.10.pdf.

Temmerman, Rita. 2000. *Towards New Ways of Terminological Description. The Sociocognitive Approach*. Amsterdam/ Philadelphia: John Benjamins.

Warnier, Maxime. 2018. “Contribution de la linguistique de corpus à la constitution de langues contrôlées pour la rédaction technique : l'exemple des exigences de projets spatiaux.” PhD diss., Université Toulouse Jean Jaurès.

Warnier, Maxime, and Anne Condamines. 2015. “A Methodology for Identifying Terms and Patterns Specific to Requirements as a Textual Genre Using Automated Tools.” In *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence*, Granada, Spain, edited by Thierry Poibeau and Pamela Faber, 183-190.

Wüster, Eugen. (1979) 1985. Introduction to the General Theory of Terminology and Terminological Lexicography. Quebec: Girsterm. Unpublished manuscript.