Lexicography ASIALEX (2018) 4:3–21
https://doi.org/10.1007/s40607-018-0034-1

ORIGINAL PAPER

Maintaining the balance between knowledge and the lexicon in terminology: a methodology based on frame semantics

Marie-Claude L'Homme[1]

Received: 28 April 2017 / Accepted: 23 January 2018 / Published online: 21 February 2018

Abstract This paper argues for an approach to terms—based on Frame Semantics (Fillmore in Ann N Y Acad Sci Conf Origin Dev Lang Speech 280:20–32, 1976; Fillmore and Baker in A Frames Approach to Semantic Analysis, 313–339, 2010)—that takes into account their linguistic properties and shows how terms and their properties are connected formally to the expression of knowledge in specialized fields. I briefly present the theoretical assumptions underlying this proposal. The main part of the article describes the methodology devised to implement the proposal in two terminological resources that are under development at the Observatoire de linguistique Sens-Texte (OLST). The methodology that comprises seven main steps is based on that of FrameNet (https://framenet.icsi.berkeley.edu/fndrupal/, 2017. Accessed 20 January 2017) (Ruppenhofer et al. in FrameNet II: extended theory and practice. https://framenet.icsi.berkeley.edu/fndrupal/index.21php?q=the_book, 2016. Accessed 27 January 2017), the lexical implementation of Frame Semantics. I illustrate the methodology by applying it to terms that belong to the field of endangered species, a subfield of the environment.

Keywords Terms · Predicative units · Frames · Frame semantics · Terminological resource · Environment

✉ Marie-Claude L'Homme
mc.lhomme@umontreal.ca

[1] Observatoire de linguistique Sens-Texte (OLST), Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada

1 Introduction

Terminology is a quite unique discipline in the sense that it must take into account knowledge as it is structured in specialized domains and observe how the lexicon (and more specifically terms) is used to express this knowledge. Theoretical proposals and/or methodologies that prevail in terminology tend to favor one perspective over the other. They can approach the problem using knowledge and its organization as a starting point and, then, link specific linguistic expressions to this knowledge. (In this paper, I will refer to this first approach as *knowledge-driven*.) Conversely, the starting point can be the lexicon, its properties and its many manifestations in texts, moving gradually to a connection with specialized knowledge. (From now on, this perspective will be called *lexicon-driven*.)

Resources compiled by terminologists can also reflect these opposite perspectives. However, in practice, most favor a knowledge-driven approach: term banks, thesauri, terminological knowledge bases and domain ontologies focus on representing or explaining knowledge in specialized fields and show how terms are connected to this knowledge. In these resources, very little information is provided on the linguistic properties of terms (the way they are used in sentences, their argument structure, collocates, etc.). In some of them, no linguistic information is provided at all. Even if there seems to be a general consensus in the scientific literature on terminology about the importance of taking into account the linguistic behavior of terms, it seems that users must still look somewhere else to obtain this kind of information.

This paper argues for an approach to terms that takes into account their linguistic properties and shows how terms and their properties are connected formally to the expression of knowledge in specialized fields. The approach is based on Frame Semantics (Fillmore 1976; Fillmore and Baker 2010), a theoretical framework devised for language in general and not specifically for terminology. I will say a few things about the approach, its theoretical assumptions and why Frame Semantics can prove interesting for terminology. We also developed a methodology based on that of FrameNet, the lexical implementation of Frame Semantics (FrameNet 2017; Ruppenhofer et al. 2016). The paper focuses mainly on the steps of this methodology and illustrates how they are applied to terms that belong to the field of endangered species.[1]

The paper is structured as follows. In Sect. 2, using some concrete examples, I will illustrate the difference between the knowledge-driven and the lexicon-driven perspectives with a series of questions raised by each one. Section 3 presents some of the reasons why Frame Semantics is an interesting framework not only to account for both knowledge and the lexicon, but also to show how these two are linked. In Sect. 4, the methodology is described and applied to terms that belong to the field of endangered species and that evoke a situation whereby species cease to exist (e.g., *disappear, extinction*). It also contains examples of the two resources; my team is

[1] I chose this field rather than one directly connected to medicine, since my team has been working on terms related to the environment and hence I am much more comfortable with these. However, the problem raised in the introduction (i.e., two different perspectives taken in terminology) remains the same regardless of the specialized domain considered.

currently compiling (the *DiCoEnviro* (2017) *Dictionnaire fondamental de l'environnement* and the *Framed DiCoEnviro* 2017) that implement our proposal.

2 Questions raised by a specialized field: endangered species

If we consider a specific field—that of endangered species—from the point of view of knowledge, lets imagine what kinds of questions may be asked:

- Name a species that is currently endangered. A possible answer: in Canada, the woodland caribou (among others).
- What is the degree of vulnerability of a given species? Species may be classified as "extinct (EX)", "extinct in the wild (EW)", "critically endangered (CR)", "endangered (EN)", "vulnerable (VU)", "near threatened (NT)", "least concerned (LC)".
- What are the causes of the decline of species? Possible answers: deforestation, human hunting, and fragmentation of habitats.
- Are there ways to prevent the disappearance of a given species?
- etc.

Some resources provide answers to this kind of questions. For instance, the Wildlife Ontology (2017) is a formal repository in which knowledge connected with species, their behavior and their conservation status is represented.

Now if we consider the field of endangered species from the point of view of the lexicon, questions formulated are likely to be very different in nature. Some examples are given below.

- What are the lexical items used to talk about endangered species? Possible answers (for English): *species, disappearance, extinct, vulnerable, abundant, survive*.
- What does vulnerable mean in the context of endangered species?
- Do some terms convey the same or similar meanings? Possible answers: *species* and *subspecies*; *disappearance, extinction*, and *extinct*; *endangered, vulnerable* and *vulnerability*.
- How does one term differ from another with regard to meaning and usage? For example, *inhabit*, *colonize* and *introduce* are related semantically since they all convey the idea of "presence in a habitat". *Inhabit* means that an animal lives in a specific habitat; *colonize* means that animals live in a habitat; however, the idea of "increasing number" is added; finally, *introduce* means that an external agent places animals in a habitat.
- How are terms related? Possible answers: *bird, lynx, polar bear; female; species; population* all denote, albeit differently, "animals"; *vulnerable; threatened; endangered* express states in which animals can find themselves; *extinction*, *disappearance* are events that can affect animals; *inhabit; reproduce, migrate* denote activities carried out by animals; *hunting, poaching* and *taking* denote activities carried out by men that can affect animals.

Terminological resources provide answers to the second series of questions to a certain extent. Many contain definitions allowing users to understand the meaning conveyed by a given term and thus help distinguish it from other terms. A large number of resources also mention synonyms, i.e., terms that denote the same concept or convey the same meaning. Terms with similar meanings can also be described indirectly, i.e., in the form of relationships established between entries or in the form of cross-references. Finally, different kinds of relationships can be represented in terminological resources. Often, however, relationships are established between the concepts denoted by terms rather than between terms themselves.

Even if they contain relevant information with respect to the second set of questions listed in this section, terminological resources tend to provide it for certain types of terms. Nouns and noun phrases are favored over verbs and adjectives; terms that denote entities are favored over events and properties. This means that although terms like *species, subspecies*, and *habitat* might be defined and perhaps related to other relevant terms; others terms such as *colonize, extinct, vulnerable*,[2] are likely to be missing. Then, possible relationships between the first terms (*species, subspecies*, etc.) and the second ones (*colonize, extinct, vulnerable*, etc.) will necessarily be lacking. And, finally, as was mentioned above, even terms that denote entities are seldom described in a way that make their linguistic properties or behavior explicit.[3]

3 Why frame semantics for terminology?

We believe that for terminological resources to be useful for all sorts of users, they should provide answers to lexicon-driven questions and, to the extent to which this is possible, to knowledge-driven ones. We assume that it is possible to connect descriptions of terms that give information about their linguistic properties to a form of representation of the knowledge they express. We approached this problem from the perspective of Frame Semantics for reasons that will become clearer in Sect. 3.2. First I will present Frame Semantics very briefly.

3.1 Frame semantics and FrameNet in a nutshell

Frame semantics (FS) (Fillmore 1976; Fillmore and Baker 2010) is a cognitive linguistics framework that is based on the assumption that the meanings of lexical units (LUs) are constructed in relation to background knowledge (built on previous experience, on beliefs, or on social conventions). Formally, the structure of this

[2] *Vulnerable* as such might not appear in terminological resources. However, the noun phrase *vulnerable species* is likely to be listed. This confirms the preference of specialized resources for nouns and noun phrases.

[3] This can easily be verified by searching for different terms via EcoRessouces (2017), an aggregator developed by the Observatoire de linguistique Sens-Texte and the research group Recherche appliquée en linguistique informatique that gives access to 16 terminological online resources containing environmental terms. If we set aside our own resource (the DiCoEnviro), all 15 others provide little linguistic information (some do not provide any kind of linguistic information).

background knowledge is represented in semantic frames. A semantic frame models a given situation; situations comprise participants, “props”, and other conceptual elements, which constitute its frame elements (FEs).

A simple example will illustrate what a semantic frame is designed to capture. Most would agree that an “eating” situation includes participants such as: (1) an animate doing the eating (a human being or an animal); and (2) something undergoing the activity (a meal, an omelet, a fruit, etc.). Other participants can also be considered: (a) an instrument used by the animate to carry out this activity (a fork, chop sticks, for instance); (b) a specific place where the eating can be done (a kitchen, a restaurant, etc.); (c) a conventional time of day during which the activity is carried out (breakfast, dinner, etc.), and so on.

In FrameNet (2017), this situation is represented in a semantic frame called **Ingestion**. It includes different kinds of participants (i.e., frame elements). Some are obligatory (called *core frame elements*): these include the Ingestor (the animate carrying out the activity) and the Ingestibles (the things being eaten). Other frame elements are optional (non-core FEs): degree, duration, instrument, means, place, etc.

In addition to giving descriptions of semantic frames that model situations, FrameNet shows how these situations are instantiated in language. In English, several lexical units “evoke” (to use Frame Semantics terminology) the **Ingestion** frame: *breakfast.v, consume.v, devour.v, dine.v, down.v, drink.v, eat.v, feast.v, feed.v, gobble.v, gulp.n, gulp.v, guzzle.v, have.v, imbibe.v, ingest.v, ingestion.n, lap.v, lunch.v, munch.v, nibble.v, nosh.v, nurse.v, put away.v, put back.v, quaff.v, sip.n, sip.v, slurp.n, slurp.v, snack.v, sup.v, swig.n, swig.v, swill.v, tuck.v.* As a native speaker of French, I can think of French lexical units that evoke the same frame, i.e., *boire,v., ingérer.v., ingestion.n, dévorer.v, manger.v., se nourir.v.*, etc. And, of course, other languages use different sets of lexical units to express the same situation.

In FrameNet, each lexical unit (LU) is described in a separate entry and appears in annotated sentences that show how LUs interact with the linguistic realizations of frame elements. Hence, lexical units and their instantiations in sentences are connected to the more abstract conceptual representation of a situation (i.e., the frame). Finally, frames can also be linked to other ones offering a wider perspective on conceptual scenarios (for example, the **Ingestion** frame is connected to the **Manipulation**, **Cause_motion**, and **Ingest_substance** frames).

3.2 Why semantic frames for terminology?

In our opinion, the assumptions about language as formulated in Frame Semantics and the modeling of frames in FrameNet are interesting for terminology for different reasons that are linked to the problem stated at the beginning of this paper (knowledge-driven vs. lexicon-driven perspectives). The potential of Frame Semantics for terminology has been recognized by different authors, and especially by Faber et al. 2016 who propose an approach called *Frame-based Terminology* “that uses certain aspects of Frame Semantics (Fillmore 1985; Fillmore and Atkins 1992) to structure specialized domains and create non-language-specific representations” (2016: 73). The methodology devised by FrameNet was also applied by terminologists to different fields of knowledge: for instance, soccer (Schmidt 2009), law (Pimentel 2013), computer science (Ghazzawi 2016).

First, part of the modeling as it is structured in FrameNet accounts for the linguistic nature of LUs. We can apply this to terms that appear in specialized corpora. For instance, annotations show that different terms that denote a situation whereby a species ceases to exist (a situation that has one obligatory participant) are used in sentences along with this obligatory participant. This is shown below with the terms *disappear*, *disappearance*, and *extinction*.

> *The likelihood that [a taxon$_{Patient}$] will* ***DISAPPEAR*** *or die out within a given area (e.g., one country or the entire world) and over a definable time period.*
>
> *Urbanisation does not only destroy biodiversity, either, even if it leads to the* ***DISAPPEARANCE*** *[of certain plants$_{Patient}$].*
>
> *The faster the rate of climate change, the greater the probability of ecosystem disruption and [species$_{Patient}$]* ***EXTINCTION****.*

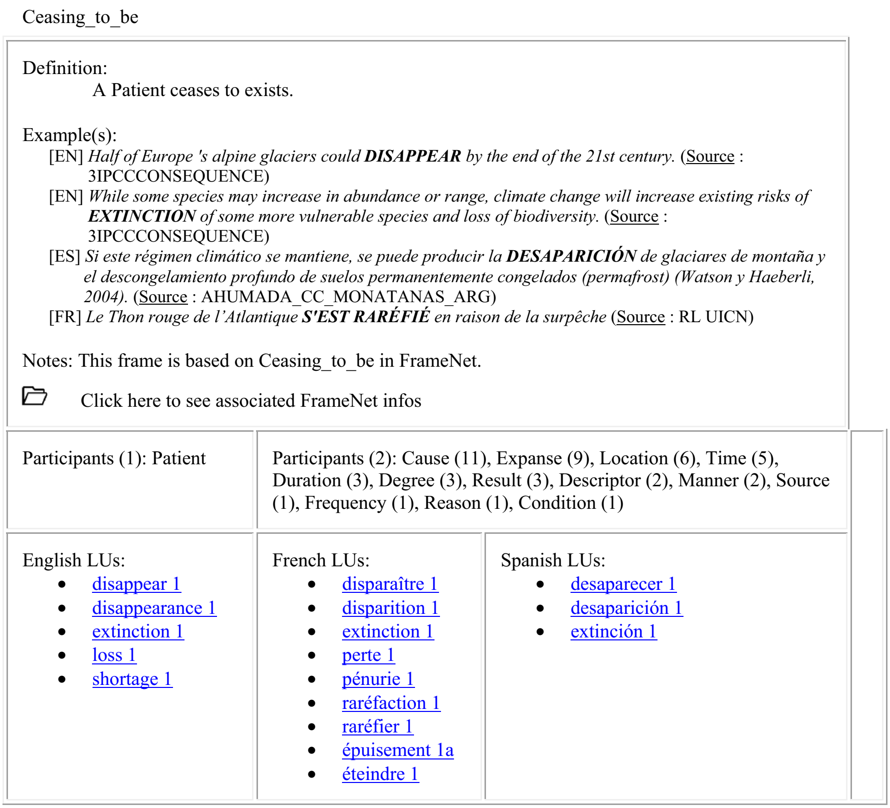
Secondly, frames provide abstract descriptions of situations evoked by lexical units. We can use frames to model situations that occur in specialized subject fields. For instance, one important situation that occurs in the field of endangered species concerns an event in which species are in the process of ceasing to exist. This frame states the obligatory participant [Participant (1): the species undergoing the process labeled as a Patient in this frame]. Obligatory participants are those that are necessary to define the situation. The frame also states the optional participants [Participants(2)] such as Cause, Location, Time, etc. These might be mentioned with reference to a situation but do not define the situation per se.[4] Terms evoking this situation in English—*disappear, disappearance, extinction*—are grouped in the same frame. Figure 1 shows how we model this situation in a frame called **Ceasing_to_be** (based on the one encoded in FrameNet) in our resource Framed DiCoEnviro (2017).

As can be seen in Fig. 1, the resource lists English terms that evoke a frame, but also French and Spanish ones.[5] Each term is described in a separate entry where lexical information and annotated contexts are provided. This information appears in a separate resource called DiCoEnviro (2017). Figure 2 is a reproduction of the entry for *extinction*.[6]

[4] Obligatory participants are labeled *core frame elements* in FrameNet and correspond partly to what is normally referred to as *arguments* (although arguments are usually defined for linguistic units; whereas frame elements are defined for frames that accounts for a conceptual representation of a situation). Optional participants are labeled *non-core frame elements* in FrameNet and correspond partially to what is called *adjuncts*.

[5] It should be noted that the terminological content of frames may be enriched as more data is taken into account. In addition, some languages are better covered that others.

[6] The DiCoEnviro is the terminological resource that contains the descriptions upon which we base our discovery of frames. It states the argument structure of terms, gives access to up to 20 contexts (when these are annotated) and lists various types of relations that terms hold with other terms in the field (related meanings, opposites, morphologically related terms, collocations, etc. In addition, when equivalents in other languages also appear in the resource (the resource covers English, French, Spanish and has some entries in Portuguese), hyperlinks are provided to allow users to access these entries. At the



Ceasing_to_be

Definition:
A Patient ceases to exists.

Example(s):
[EN] *Half of Europe 's alpine glaciers could **DISAPPEAR** by the end of the 21st century.* (Source : 3IPCCCONSEQUENCE)
[EN] *While some species may increase in abundance or range, climate change will increase existing risks of **EXTINCTION** of some more vulnerable species and loss of biodiversity.* (Source : 3IPCCCONSEQUENCE)
[ES] *Si este régimen climático se mantiene, se puede producir la **DESAPARICIÓN** de glaciares de montaña y el descongelamiento profundo de suelos permanentemente congelados (permafrost) (Watson y Haeberli, 2004).* (Source : AHUMADA_CC_MONATANAS_ARG)
[FR] *Le Thon rouge de l'Atlantique **S'EST RARÉFIÉ** en raison de la surpêche* (Source : RL UICN)

Notes: This frame is based on Ceasing_to_be in FrameNet.

🗁 Click here to see associated FrameNet infos

Participants (1): Patient	Participants (2): Cause (11), Expanse (9), Location (6), Time (5), Duration (3), Degree (3), Result (3), Descriptor (2), Manner (2), Source (1), Frequency (1), Reason (1), Condition (1)

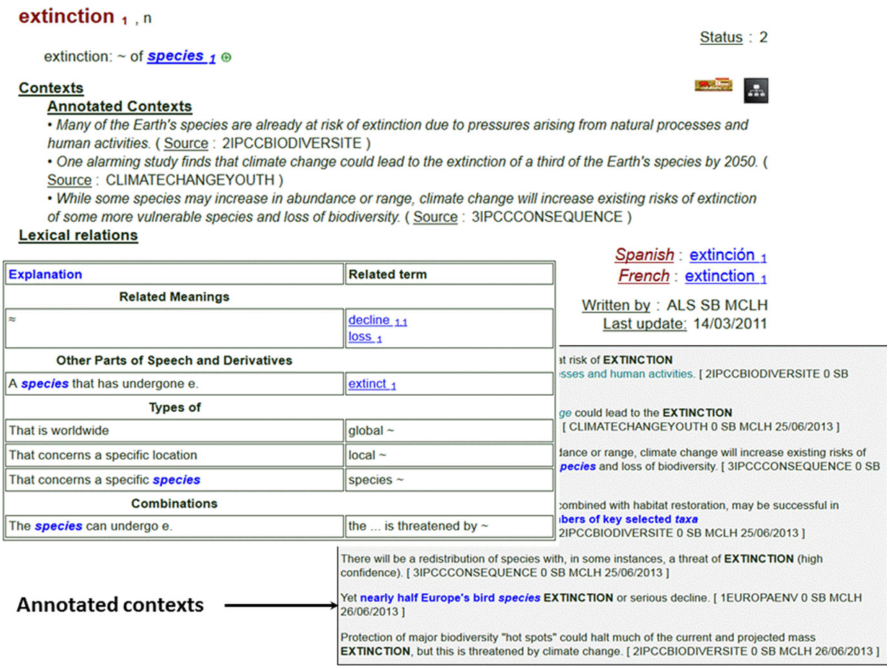
English LUs:	French LUs:	Spanish LUs:
• disappear 1 • disappearance 1 • extinction 1 • loss 1 • shortage 1	• disparaître 1 • disparition 1 • extinction 1 • perte 1 • pénurie 1 • raréfaction 1 • raréfier 1 • épuisement 1a • éteindre 1	• desaparecer 1 • desaparición 1 • extinción 1

Fig. 1 The frame **Ceasing_to_be** in the Framed DiCoEnviro (2017)

Finally, as in FrameNet, frames can be linked to give a wider perspective on activities that occur in the field. Figure 3 shows how the **Ceasing_to_be** frame is connected to other frames in the field of endangered species. Relations established between frames in the Framed DiCoEnviro (2017) are often domain specific. For instance, the **Ceasing_to_be** frame is preceded by **Surviving** that comprises terms such as *persist*, *survive*, and *survival*.

Frames, the way we define them, relations between frames, and contextual annotations are further explained in Sects. 4.5, 4.6 and 4.7.

Footnote 6 continued
end of July 2017, the DiCoEnviro contained 884 English entries (with over 4000 lexical relations and 8000 annotated contexts) and 1264 French entries (with over 6500 lexical relations and 20,000 annotated contexts). The resource also includes a few Spanish and Portuguese terms. The DiCoEnviro is first designed as a tool for researchers in terminology, but some of the information it contains (the annotated contexts, lexical relations) makes it attractive to other kinds of users, i.e., translators, lexicographers, etc.



extinction 1 , n

Status : 2

extinction: ~ of ***species*** 1

Contexts

Annotated Contexts

- *Many of the Earth's species are already at risk of extinction due to pressures arising from natural processes and human activities.* (Source : 2IPCCBIODIVERSITE)
- *One alarming study finds that climate change could lead to the extinction of a third of the Earth's species by 2050.* (Source : CLIMATECHANGEYOUTH)
- *While some species may increase in abundance or range, climate change will increase existing risks of extinction of some more vulnerable species and loss of biodiversity.* (Source : 3IPCCCONSEQUENCE)

Lexical relations

Explanation	Related term
Related Meanings	
≈	decline 1.1 loss 1
Other Parts of Speech and Derivatives	
A ***species*** that has undergone e.	extinct 1
Types of	
That is worldwide	global ~
That concerns a specific location	local ~
That concerns a specific ***species***	species ~
Combinations	
The ***species*** can undergo e.	the ... is threatened by ~

Spanish : extinción 1
French : extinction 1

Written by : ALS SB MCLH
Last update: 14/03/2011

at risk of **EXTINCTION**
sses and human activities. [2IPCCBIODIVERSITE 0 SB

ge could lead to the **EXTINCTION**
[CLIMATECHANGEYOUTH 0 SB MCLH 25/06/2013]

dance or range, climate change will increase existing risks of
pecies and loss of biodiversity. [3IPCCCONSEQUENCE 0 SB

combined with habitat restoration, may be successful in
bers of key selected *taxa*
2IPCCBIODIVERSITE 0 SB MCLH 25/06/2013]

There will be a redistribution of species with, in some instances, a threat of **EXTINCTION** (high confidence). [3IPCCCONSEQUENCE 0 SB MCLH 25/06/2013]

Yet **nearly half Europe's bird *species*** **EXTINCTION** or serious decline. [1EUROPAENV 0 SB MCLH 26/06/2013]

Protection of major biodiversity "hot spots" could halt much of the current and projected mass **EXTINCTION**, but this is threatened by climate change. [2IPCCBIODIVERSITE 0 SB MCLH 26/06/2013]

Fig. 2 Entry *extinction* in the DiCoEnviro (2017)

4 A seven steps methodology

To enrich the resources presented in Sect. 3.2, we developed a methodology based on that devised within the FrameNet project (Fillmore et al. 2003; Ruppenhofer et al. 2016) for general English.

Some adaptations were made (L'Homme 2015, 2016) to account for the specialized nature of the lexicon with which we are dealing and our own objectives. The methodology is bottom-up[7] and combines automated tools and manual analysis. It comprises seven main steps that are described below using terms that evoke the same frame.

4.1 Compilation of specialized corpora

The first step of most terminological projects including ours consists in compiling a corpus. Very few available specialized corpora are available[8] and terminologists

[7] A bottom-up methodology was also used by other researchers interested in specialized lexica (Pimentel 2013; Schmidt 2009).

[8] There are exceptions though. In the field of the environment, for instance, a corpus called PANACEA (2015) can be used for research purposes. However, the corpus was compiled automatically and might not be suitable for our terminological projects since automatically compiled corpora do not discriminate textual genres dealing with the same topic (scientific articles, reports, newspaper articles). Some even contain glossaries that do not show how terms are used in running texts. Since we want to be able to know

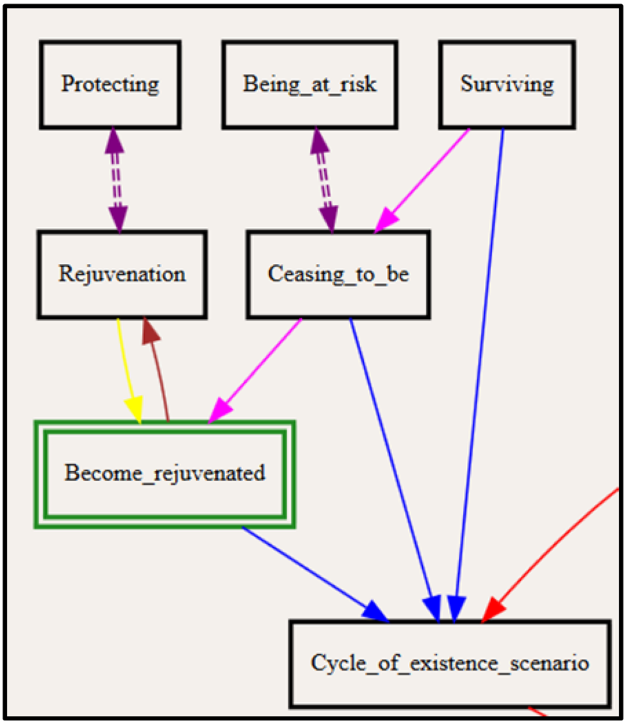


Fig. 3 **Ceasing_to_be** and related frames in the Framed DiCoEnviro (2017)

must devote part of their time locating relevant texts that will placed in a corpus used in later steps of the methodology.

Corpora are used in all lexicographical projects, but they are especially important in terminology. Terminologists are seldom experts of the field they must describe and they rely heavily on the contents of corpora to locate relevant terms and information about them. They cannot rely on intuition as much as lexicographers. Hence, corpora are viewed as the main repositories of knowledge and the information found in the corpus can be validated by an expert (at least for problematic cases).

Since the field of the environment encompasses many different topics (renewable energy, sustainable development, climate change, etc.) and that the terminology and the number of occurrences of given terms can vary quite drastically from one topic to another, we identify specific topics and compile a corpus accordingly (and do this for the different languages we take into account). Up to now, we compiled seven different English corpora that deal with climate change, renewable energy, endangered species, deforestation, waste management, electric vehicles, and water pollution. Of course, other topics will be covered in the future.

Footnote 8 continued
exactly where contexts that we collect come from and record many details regarding texts that we place into our corpora, compiling corpora manually still remains the best option.

For the time being, our method is rather simple and consists in locating relevant texts on the web with specific key words. For endangered species, the name of the field can be entered in a search engine (“endangered species”) with or without other key words (protection, animals, birds, bees, habitat, etc.). Documents returned by the search engine are selected if they meet criteria such as authoritative source, length, direct relationship with the topic, etc. They are then converted into raw text. When we embark on a new project, we start with corpora of about 500,000 words (this corresponds to 30-40 different texts of varying sizes, between 1000 and 50,000 occurrences). Corpora are often enriched at a later stage (for instance, our English corpus on endangered species now amounts to 1,058,869 occurrences and comprises 88 different texts).

4.2 Identification of terms

Once a corpus on a specific topic is compiled, we proceed to identify relevant terms. We first approach this task with an automated method that produces a list of candidate-terms. Then the list is filtered by a terminologist.

We submit our corpus to a term extractor, called *TermoStat* and developed by Drouin (2003). The term extractor automatically compares the content of our specialized corpus to a reference corpus. For English, the reference corpus is a combination of the British National Corpus (BNC) and the American National Corpus (ANC). The extractor compares lemmatized and part of speech tagged units in both corpora and produces a list of candidate-terms according to their specificity in the specialized corpus. This specificity is a reflection of the unusual frequency of the unit in the specialized corpus. The hypothesis underlying this method is that unusually frequent units correspond to terms. Table 1 shows the first results of this method applied to our corpus of endangered species.

Terminologists must then analyze this list, keep those candidates that correspond to relevant terms, and ignore other lexical items. Although some cases do not raise problems (e.g., *species*, *habitat*), others might be much more problematic. For these cases, four criteria can be applied.

For instance, the unit *extinction* was extracted by the automated method. Terminologists may determine its terminological nature by looking at contexts in which the unit appears (a sample is given below) and analyzing the types of arguments it combines with. If the arguments are realized in the form of terms, then the unit is likely to be a term itself.

> *The faster the rate of climate change, the greater the probability of ecosystem disruption and* species ***extinction***.

> *While some species may increase in abundance or range, climate change will increase existing risks of* ***extinction*** of some more vulnerable species *and loss of biodiversity.*

Captive breeding and translocation, when combined with habitat restoration, may be successful in preventing the ***extinction*** of small numbers of key selected taxa *under small to moderate climate change.*

Table 1 First term candidates extracted from a corpus on endangered species

Canonical form	Frequency	Specificity score	Variants
specie	3710	202.96	specie, species
species	3046	185.74	species
habitat	2614	173.96	habitat, habitats
conservation	1388	112.76	conservation
recovery	1142	108.22	recovery, recoveries
endangered	928	103.71	endangered
population	1621	98.61	population, populations
threaten	943	84.46	threaten, threatens, threatened, threatening
extinction	603	81.63	extinction, extinctions
endanger	504	72.48	endanger, endangered, endangering
status	866	71.86	status
nest	422	69.94	nest, nests, nested, nesting
threat	789	67.43	threat, threats

Variants correspond to inflected forms identified by a tool used to accomplish this task (TreeTagger (Schmid 1994)) before term extraction is performed. The statistical measures applied by TermoStat is based on lemmas and not on inflected forms

Since lemmatization is automated, it might produce some erroneous entries

These contexts show that the arguments of *extinction* are realized as *species* or *taxa*, which are two relevant terms in the field of endangered species. Terminologists will thus keep *extinction* as a relevant term as well and add it to the terminological resource DiCoEnviro (2017). It is important to say at this point that term candidates appear in two different entries, since some are polysemic. For instance, *environment* conveys two different meanings in the domain: (1) “The set of conditions, influences that characterizes the Earth of one of its components” (*impacts on the environment, protection of the environment*); (2) “A set of factors—as climate, soil, and living things—that acts on an organism or an ecological community and determines its form and survival” (*organisms and their physical environment; a protected species and its environment*). Each meaning is described in a separate entry.

4.3 Extraction of contexts

The third step of our methodology consists in going back to the corpus (our corpus on endangered species) and retrieving contexts that will be placed in the entry. These contexts are extremely useful to analyze the term and complete parts of its description. They are also annotated, as will be seen in Sect. 4.5.

For each term, terminologists extract 15–20 different contexts. These are selected according to the richness of the information they contain (presence of participants, explanations of the meaning, etc.) (a sample is given below for the term *extinction*). Experience has shown that 15–20 contexts are sufficient to give a clear picture of how terms behave in a specialized corpus.

Extinction (selected as a relevant term)

Many of the Earth's species are already at risk of ***EXTINCTION*** *due to pressures arising from natural processes and human activities. [2IPCCBIODIVERSITE]*

One alarming study finds that climate change could lead to the ***EXTINCTION*** *of a third of the Earth's species by 2050. [CLIMATECHANGEYOUTH]*

While some species may increase in abundance or range, climate change will increase existing risks of ***EXTINCTION*** *of some more vulnerable species and loss of biodiversity. [3IPCCCONSEQUENCE]*

Captive breeding and translocation, when combined with habitat restoration, may be successful in preventing the ***EXTINCTION*** *of small numbers of key selected taxa under small to moderate climate change. [2IPCCBIODIVERSITE]*

There will be a redistribution of species with, in some instances, a threat of ***EXTINCTION*** *(high confidence). [3IPCCCONSEQUENCE]*

At this stage, terminologists might make meaning distinctions that they missed during the previous step. Since distinct meanings are described in separate entries, contexts must reflect these distinctions and be placed in the right entry.

4.4 Definition of the argument structure

The fourth step consists in defining the argument structure of terms. This step—albeit central in our methodology—does not apply to terms that are non-predicative (e.g., *animal, organism, plant, and wolf*). At this stage, terminologists determine how many arguments a term has and state these arguments in the entry. For instance, *extinction* as one argument: *extinction* of X.

We use two different systems to represent arguments in the terminological resource DiCoEnviro (see Fig. 2). We first label them with semantic roles that must express the relationship between the term and its arguments. The label used for the argument of *extinction* is Patient[9] (in Fig. 2, this can be seen by the use of the color blue chosen to represent Patients); this role expresses the fact that the argument is the one undergoing the "extinction". An additional label states what we call a *typical term*. This typical term is designed to give the user an idea of the kinds of terms that can instantiate this argument. In the argument structure of *extinction*, the typical term *species* was chosen.

4.5 Annotation of contexts

Once the argument structure is defined, terminologists proceed to annotate the 15–20 contexts retrieved for terms based on the methodology devised for the

[9] It should be said at this point that labels used in our terminological resources differ from those used in FrameNet. Frame elements in FrameNet are relevant within a specific frame. In our resources, labels should be applied to large sets of terms.

FrameNet project (Ruppenhofer et al. 2016). The annotation consists in making explicit the following items in each context:

- The target: the term itself (e.g., *extinction*);
- The participants of the target: the realizations of the arguments; we also annotate adjuncts);
- The semantic roles of the participants: Agent, Patient, Cause, etc.
- The syntactic functions of participants: subject, modifier, etc.
- The syntactic groups of participants: NP, PP, etc.

The examples below are a sample of the annotated contexts for the term *extinction*. The graphical display appears in Fig. 2 (the table below the contexts is a summary of the properties of participants).

> [Many of the Earth's species$_{\text{Patient}}$] are already at risk of **EXTINCTION** due to pressures arising from natural processes and human activities. [2IPCCBIODIVERSITE 0 SB MCLH 25/06/2013]
>
> One alarming study finds that climate change could lead to the **EXTINCTION** [of a third of the Earth's species$_{\text{Patient}}$] by 2050. [CLIMATECHANGEYOUTH 0 SB MCLH 25/06/2013]
>
> While some species may increase in abundance or range, climate change will increase existing risks of **EXTINCTION** [of some more vulnerable species$_{\text{Patient}}$] and loss of biodiversity. [3IPCCCONSEQUENCE 0 SB MCLH 25/06/2013]
>
> Captive breeding and translocation, when combined with habitat restoration, may be successful in preventing the **EXTINCTION** [of small numbers of key selected taxa$_{\text{Patient}}$] [under small to moderate climate change$_{\text{Condition}}$]. [2IPCCBIODIVERSITE 0 SB MCLH 25/06/2013]
>
> Yet [nearly half Europe's bird species$_{\text{Patient}}$] **EXTINCTION** or serious decline. [1EUROPAENV 0 SB MCLH 26/06/2013]

Extinction 1		
Arguments		
Patient	Complement (PP-of) (3) Indirect link (NP) (2) Modifier (NP) (1)	species (4) taxon (1)
Others		
Condition	Complement (PP-under)	change

4.6 Definition of semantic frames

Once the argument structure is defined (step 4.4) and contexts annotated (step 4.5), terminologists proceed to identify terms that are likely to evoke the same frame. This identification is guided by different lexico-semantic properties of terms that are described in their entries:

- The same number of arguments: e.g., *extinction* and *disappear* both have one argument;
- Arguments are of a similar nature: e.g., the arguments of *extinction* and *disappear* are labeled as Patients, and are instantiated by terms that denote living organisms and/or habitats (*animal, ecosystem, fish, plant, species*, etc.)
- Shared adjuncts: e.g., adjuncts expressing the Expanse appear in contexts for *extinction* and *disappear* (*globally, completely, within a given area*).

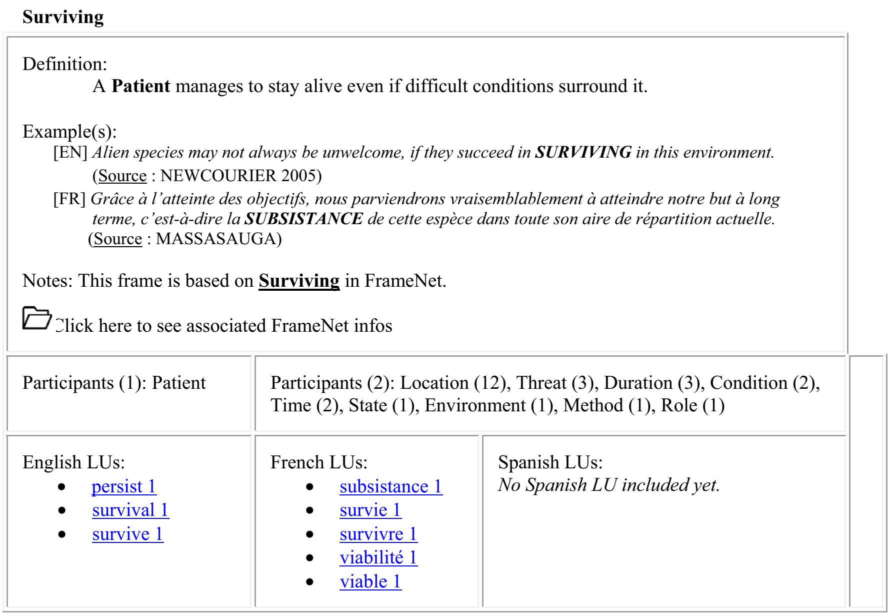
Of course, shared participants are useful clues to identify terms evoking the same frame, but terminologists must define the content of frames based on much more. Terms must denote the same general situation and share presuppositions about it. Hence, based on our descriptions, we were able to establish that the terms *extinction*, *disappear* and *disappearance* evoke the same situation, whereby a living entity is in the process of ceasing to exist. This process can be further characterized as affecting a small or large number of these entities and might be caused by specific conditions.[10]

To help them define frames, terminologists refer to FrameNet. They try to find corresponding data in the English: more specifically, they try to find lexical units in FrameNet that correspond to terms that appear in the terminological resource. If a frame was already encoded in FrameNet and that the data it describes fits the properties of terms in the field of endangered species, the frame as defined in FrameNet is used and adapted. For instance, this was possible for the terms *disappear* and *disappearance*. They both appear in FrameNet and the situation they evoke is the same in the field of the environment. We thus based our frame on that of FrameNet. Of course, many differences appear in the descriptions given in each resource (different lexical content, labels used for participants, etc.). Furthermore, when we base our frame on an existing one in FrameNet, we use the same name and provide a link that will lead users to its description in original the FrameNet resource.[11]

There are many cases for which no correspondence can be established and we must create frames that account for our specific data. For instance, we created a frame called **Adding_trees_in_location** for terms such as *afforest*, *reforest*, *afforestation*, *reforestation*. More than half of the frames that appear in the Framed

[10] A member of our team (Bernier-Colborne 2016) explored how a method based on distributional semantics to identify the terminological content of frames automatically. The method is promising but has not been completely integrated to our methodology.

[11] Users can also view the similarities and the differences between frames as they are represented in FrameNet and those that appear in the Framed DiCoEnviro when selecting the “Click here to see associated FrameNet infos”. More explanations are given about this in L’Homme et al. (2016).



Surviving

Definition:
A **Patient** manages to stay alive even if difficult conditions surround it.

Example(s):
[EN] *Alien species may not always be unwelcome, if they succeed in* ***SURVIVING*** *in this environment.*
(Source : NEWCOURIER 2005)
[FR] *Grâce à l'atteinte des objectifs, nous parviendrons vraisemblablement à atteindre notre but à long terme, c'est-à-dire la* ***SUBSISTANCE*** *de cette espèce dans toute son aire de répartition actuelle.*
(Source : MASSASAUGA)

Notes: This frame is based on **Surviving** in FrameNet.

Click here to see associated FrameNet infos

Participants (1): Patient	Participants (2): Location (12), Threat (3), Duration (3), Condition (2), Time (2), State (1), Environment (1), Method (1), Role (1)	
English LUs: • persist 1 • survival 1 • survive 1	French LUs: • subsistance 1 • survie 1 • survivre 1 • viabilité 1 • viable 1	Spanish LUs: *No Spanish LU included yet.*

Fig. 4 The frame **Surviving** in the Framed DiCoEnviro (2017)

DiCoEnviro (2017) were defined specifically to account for situations in the field of the environment. In these cases, we create a name that should be evocative of the situation that it represents.

Once frames are defined, they are encoded in an entry that accounts for the following (see Fig. 4 for the frame **Surviving**; another example was given above for the frame **Ceasing_to_be**, Fig. 1):

- The name of the frame: **Surviving**;
- A definition formulated for the field of the environment and stating the obligatory participants;
- Example(s) for each of the languages described;
- An indication of the reference to FrameNet with a hyperlink to FrameNet wherever relevant;
- The participants: (1) obligatory; (2) optional;
- The list of terms that evoke this frame; a hyperlink to the DiCoEnviro (2017) is provided to visualize the terminological entry and contextual annotations.

4.7 Defining relations between frames

Situations are connected in different ways and frames that capture these situations can be linked so as to make these connections explicit. For instance, the **Ceasing_to_be** frame in the field on endangered species is linked to **Surviving** (Fig. 4), since before they cease to be, species find different ways of surviving. It is

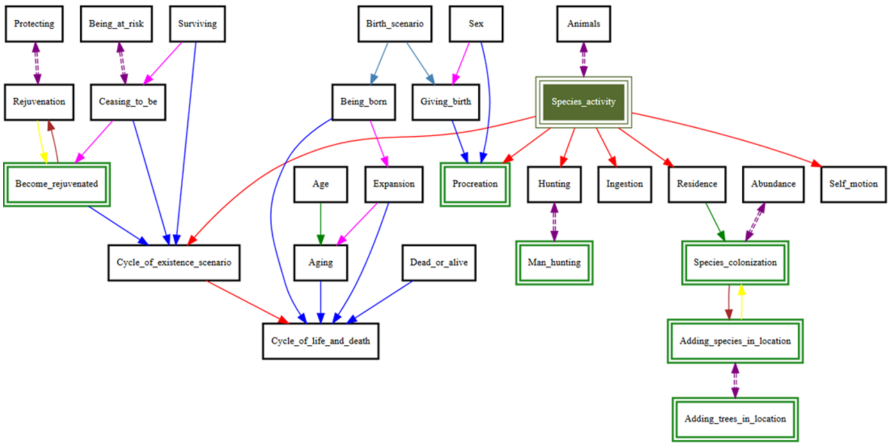


Fig. 5 The **Species_activities_and_life** scenario in the Framed DiCoEnviro (2017)

also linked to another frame called **Rejuvenation** that captures a situation whereby species can recover. Finally, another frame that captures a situation in which species are in a state where they are exposed to or otherwise liable to be affected by a Threat and called **Being_at_risk** is also linked to **Ceasing_to_be**. The relationships mentioned in this paragraph are presented above in Fig. 3.

In the Framed DiCoEnviro (2017), frames are linked through different kinds of relationships that are listed below (some of these relationships appear in Fig. 5 and are identified with different colors). The first seven are based on those defined in FrameNet; the last two were defined for the frames in our resource.

- Inheritance: Inherits, Is inherited by (e.g., **Cycle_of_life_and_death** inherits **Cycle_of_existence scenario**);
- Perspective: Perspective on, Is perspectivized in (e.g., **Giving_birth** and **Being_born** perspective in a different way **Birth_scenario**);
- Use: Uses, Is used by (e.g., **Species_colonization** Uses **Residence**);
- Subframe: Subframe of, Has subframe (e.g., **Giving_birth** Subframe of **Birth_scenario**);
- Precedence: Precedes; Is preceded by (e.g., **Being_born** Precedes **Expension**);
- Causation: Is inchoative of; Is causative of (e.g., **Species_colonization** Is inchoative of **Adding_species_in_location**);
- See also: See also (e.g., **Being_at_risk** See also **Ceasing_to_be**).
- Opposition: Is opposed to (e.g., **Removing_trees_from_location** Is opposed to **Adding_trees_in_location**).
- Property: Is a property of, Has property (e.g., **Sustainability** is a property of **Human_activity**).

Once linked, frames can lead to larger scenarios that give a first overview of how events are connected in the field of the environment. For instance, the scenario reproduced in Fig. 5 shows the different activities that species undergo or carry out.

Another scenario, called *Understanding life*, shows the different connections between living organisms according to the terms used to express them (*species, population, predator, offspring*, etc.).

5 Concluding remarks

In this article I presented an approach and its associated methodology designed to take into account the linguistic properties of terms and connect these terms and their properties to a structure that could correspond to the representation of knowledge in a given specialized field. The entries found in the terminological resource DiCoEnviro (2017) supply information about linguistic properties with the argument structure and contextual annotations. Then these entries are connected to another resource, called the Framed DiCoEnviro (2017) and are linked to frames that they evoke. Then frames are grouped into scenarios designed to provide a more global picture of the different events that occur in the field of the environment or the different entities that cause or undergo these events. This approach is based on Frame Semantics and adapted from the methodology developed within the FrameNet project. We showed, among other things, that our approach can take into account types of terms that are often ignored in terminology, namely terms that denote events and activities. We believe that our methodology can be easily adapted to other languages and other fields of knowledge. It was recently applied in Ghazzawi (2016) to the terminology of computer science for Arabic, English and French terms. However, although some members of the team explored different ways to automate parts of the methodology (Azoulay 2017; Bernier-Colborne 2016; Forest et al. 2015; Hadouche et al. 2011), we are aware that more can be done in this area. In addition to alleviating time-consuming work, increased automation could make the method more attractive to researchers interested in modeling terms in other fields of knowledge.

Although our methodology is rather stable and has been applied to different corpora and a fairly large number of terms, the two resources to which I referred in this article are under construction. We are in the process of adding lexical content to frames, defining new frames, and establishing relations between frames. Our bottom-up methodology allows us to enrich the contents of our resources as we discover new evidence in the data. For the time being, the Framed DiCoEnviro (2017) includes the descriptions of 184 different frames. A total of 413 English terms, 595 French terms and 45 Spanish terms are distributed throughout these 184 frames (including 15–20 annotated contexts for each term).

We also developed 16 larger scenarios that give a general overview of important aspects of the field of the environment: e.g., Understanding the Earth, Understanding life, Substances and resources, Changes affecting the environment, Species activities and life (in which the **Cease_to_be** and **Surviving** frames appear), Human activities and their impact on the environment. Users can start browsing a general scenario, then focus on a specific frame, and finally access the terms that evoke this frame and their linguistic properties.

It is difficult, if not impossible, at this stage to evaluate how many frames and scenarios will be necessary to account for all the situations that occur in the field of the environment. The field encompasses so many topics that such an estimation would be extremely risky. However, we are confident that our methodology is sound enough to be able to account for many more frames and scenarios.

Acknowledgements This research is supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada and by the Fonds de recherche du Québec—Société et culture (FRQ-SC). I would like to thank the members of my research team who contributed in one way or another to the Framed DiCoEnviro project. I also extend my thanks to two anonymous reviewers whose comments helped clarify many parts of the paper.

References

Azoulay, D. 2017. Frame-based knowledge representation using large specialized corpora. In: *Proceedings of the AAAI spring symposium on computational construction grammar and natural language understanding*, Stanford University, CA.

Bernier-Colborne, Gabriel. 2016. *Aide à l'identification de relations lexicales au moyen de la sémantique distributionnelle et son application à un corpus bilingue du domaine de l'environnement*. Ph.D Thesis presented at the Université de Montréal, Montréal.

DiCoEnviro. 2017. *Dictionnaire fondamental de l'environnement*. http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi. Accessed 31 July 2017.

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

EcoRessources. *Terminological resources for the environment*. 2017. http://termeco.info/EcoRessources/index-e.html. Accessed 31 July 2017.

Faber, P., P. León-Araúz, and A. Reimerink. 2016. EcoLexicon: New features and challenges. In *GLOBALEX 2016: lexicographic resources for human language technology and 10th edition of the language resources and evaluation conference*, ed. by Kernerman, I., I. Kosem Trojina, S. Krek, and L. Trap-Jensen, 73-80. Portorož.

Fillmore, C.J. 1976. Frame semantics and the nature of language. *In Annals New York Academy of Sciences: Conference on the Origin and Development of Language and Speech* 280: 20–32.

Fillmore, C.J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6: 222–254.

Fillmore, C. J., and B.T. Atkins. 1992. Toward a frame-based Lexicon: the semantics of RISK and its neighbors." In Frames, Fields and Contrasts, ed. by A. Lehrer, and E. Feder Kittay, 75–102. Hillsdale, New Jersey: Lawrence Erlbaum Assoc.

Fillmore, C.J., and C. Baker. 2010. A frames approach to semantic analysis. In *Handbook of Linguistic Analysis*, ed. B. Heine, and H. Narrog, 313–339. Oxford: Oxford University Press.

Fillmore, C., M.R.L. Petruck, J. Roppenhofer, and A. Wright. 2003. FrameNet in action: the case of attaching. *International Journal of Lexicography* 16 (2): 297–332.

Forest, D., H. Brousseau, P. Drouin, and G. Bernier-Colborne. 2015. L'environnement vu par ses documents: utilisation de techniques de fouille de textes dans un contexte de description linguistique. In *13e Journées internationales d'analyse statistique des données textuelles*, Nice, France.

Framed DiCoEnviro. 2017. *A Framed Version of DiCoEnviro*. http://olst.ling.umontreal.ca/dicoenviro/framed/index.php. Accessed 31 July 2017.

FrameNet. 2017. https://framenet.icsi.berkeley.edu/fndrupal/home. Accessed 20 January 2017.

Ghazzawi, N. 2016. *Du terme prédicatif au cadre sémantique: méthodologie de compilation d'une ressource terminologique pour les termes arabes de l'informatique*. Ph.D. Thesis, presented at the Université de Montréal, Montreal.

Hadouche, F., S. Desgroseillers, J. Pimentel, M.C. L'Homme, and G. Lapalme. 2011. Identification des participants de lexies prédicatives: évaluation en performance et en temps d'un système d'annotation automatique. In *Terminologie et intelligence artificielle* (TIA 2011), Institut national des langues orientales INALCO, Paris.

L'Homme, M.C. 2015. Découverte de cadres sémantiques dans le domaine de l'environnement: le cas de l'influence objective. *Terminàlia* 12: 29–40.

L'Homme, M.C. 2016. Terminologie de l'environnement et sémantique des cadres. In *Congrès mondial de linguistique française* (CMLF 2016), Tours, France.

L'Homme, M.C., C. Subirats, and B. Robichaud. 2016. A Proposal for combining "general" and specialized frames. In *Proceedings of the workshop on cognitive aspects of the Lexicon*. 156–165, Osaka, Japan.

PANACEA. 2015. http://panacea-lr.eu/en/info-for-researchers/data-sets/monolingual-corpora. Accessed 23 January 2017.

Pimentel, J. 2013. Methodological bases for assigning terminological equivalents. A Contribution. *Terminology* 19 (2): 237–257.

Ruppenhofer, J, M. Ellsworth, M. Petruck, C. Johnson, and C. Baker, and J. Scheffczyk. 2016. *FrameNet II: extended theory and practice*. https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=the_book. Accessed 27 January 2017.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, Manchester, UK.

Schmidt, T. 2009. The Kicktionary—a multilingual lexical resources of football language. In *Multilingual FrameNets in Computational Lexicography. Methods and Applications*, ed. Boas, H.C., 101–134. Berlin/New York: Mouton de Gruyter.

Wildlife Ontology (2017). http://www.bbc.co.uk/ontologies/wo. Accessed 20 January 2017.