Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50.
DOI 10.1075/term.21.1.02sep. ISSN 0929–9971 / E-ISSN 1569–9994 

# An Ontological Framework for Modeling the Contents of Definitions

Selja Seppälä

This paper addresses the troublesome question of feature selection and content prediction in definition writing. I present the basis of definition-authoring tools that can be used across a range of contexts, independently of the domain and language of the definitions. In addition to being domain- and language-independent, these tools should be easily tailorable to specific domains. Thus, my work seeks to contribute to developing generic definition-writing aids that can be tailored to a range of different contexts and domains. The objectives of this article are: (1) to show that it is possible to create implementable *generic* definition models; (2) to show how to constrain these models to produce definitions *relevant* to particular contexts; and (3) to propose an *ontological analysis framework* with a fixed and well-motivated descriptive vocabulary that can be used in further content analysis studies in terminology and to enhance integration of textual definitions in ontologies.

Keywords: definitions, definition contents, the Basic Formal Ontology (BFO), ontologies, feature relevance, ontological analysis, conceptual analysis, ontological annotation schema, metalanguage

## 1 Introduction

Definitions are essential in terminological dictionaries such as *TERMIUM Plus*® (2014) and *Le grand dictionnaire terminologique* (GDT, 2014). They are also central to terminological databases used in public and private organizations, such as *The terminology database of the Swiss Federal Administration* (TERMDAT, 2014) and the *UBS Dictionary of Banking* (2014). Such resources include definitions in order to convey information about the meanings of domain-specific terms in one or more languages. The definitions help users to learn new terminologies, and they facilitate and enhance communication. In ontologies such as the *Gene Ontology* (GO, 2014) and the *Infectious Disease Ontology* (IDO, 2014), natural language definitions are important to help biomedical scientists, including curators of experimental literature, to correctly understand the intended meanings of ontology terms and thus avoid errors when such terms are used in the annotation (semantic tagging) of biomedical texts.

Despite the centrality of definitions, the activity of creating definitions must still be realized manually; this is time-consuming, costly, requires uncommon expertise, and is prone to all kinds of inconsistencies. Terminologists have access to lists of conventional rules and to publishers' specifications regarding the forms definitions must take; but these relate primarily to issues such as punctuation and capitalization which can be automatically checked (Seppälä 2006a, 2006b). However, regarding the *content* of definitions, they rarely have access to more than general and vague principles; software tools that might help them in writing definitions do not, as yet, exist. Such tools could accelerate the writing process and increase the consistency of definitions. This would bring further advantages, such as enhanced reusability of terminological dictionaries and databases as lexical resources, for example, in natural language processing (NLP) and for domain-ontology development, a field that has growing ties with terminology research (Grabar et al. 2012; Aussenac-Gilles et al. 2008).

Here, I present the basis for creating definition-authoring tools that can be used across a range of contexts, independently of the domain and language of the definitions. In addition to being (as far as possible) domain- and language-independent, the tools would also be easily adaptable to specific domains. Thus, my

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

work seeks to contribute to developing generic definition-writing aids (primarily software, but also principles and methods) that can be tailored to particular contexts or domains.

## 2 Background

The task of developing domain- and language-independent definition-writing tools calls for a consistent and integrated understanding of what terminological (specialized) definitions are and of what determines their contents.

Definition-writing manuals offer rather general statements focusing mainly on the logical and linguistic forms of definitions — e.g., *systematicity*, *simplicity*, *affirmativeness* — and on errors to avoid, such as *circularity* or *tautology* (Dubuc 1978; ISO 704 2009; Pavel and Nolet 2001; Rondeau 1984; Rousseau 1983; Suonuuti 1997; Vézina et al. 2009). These manuals are not, in the main, intended to establish a theory of definitions. However, even where one finds the beginnings of a theoretical approach, many questions remain unclear. The scholarly literature in terminology mostly consists of typologies of definitions (Blanchon 1997; Clas 1985; Larivière 1996; Pozzi 2001; Sager 1990), of methodological descriptions of the sort we find in practical manuals (Célestin et al. 1988; Dubuc 1978; Rondeau 1984), and of short accounts in references of a broader scope (Cabré 1998; Picht and Draskau 1985; Sager 1990; Temmerman 2000a). More substantial works focus on representational and methodological issues (Faber 2002; Faber Benítez 2009; García de Quesada 2001; García de Quesada et al. 2002), but tend to focus on only a small selection of the many possible factors that may influence the contents of definitions (see section 3 below).

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994

## 2.1 Terminological Definitions

We start with Marconi's theory of *lexical competence* (Marconi 1997), according to which the meaning of a term (a sign's semantic value) arises when competent speakers — e.g., the experts in a given domain — consistently use a string of letters or symbols to communicate about something in the world.[5] This something is the referent of the term — for example the objects, processes, or attributes relevant to the experts' activity.

Meaning is then something like a mental representation (or concept) that experts have in mind when using a term to refer either (1) to a type of thing (for example, *particle accelerator*) or (2) to a particular thing (for example, the *Large Hadron Collider*) in the world as delimited within their domain of expertise. In both cases, the mental representation expressed by a term is composed of certain features that are (at least partly) also expressible in a definition.

Thus, the term, the mental representation, and the definition all have the same referent. Considering that this referent is the object of the definition (that which the definition is about), the expressions *to define a term*, *to define a concept* and *to define a thing* can be used synonymously, as will be the case in what follows.

A *terminological definition* is a short piece of text found in the *definition field* of a specialized dictionary entry, and in the *definition* or *comment* fields of an ontology document. Roughly speaking, it corresponds to a single sentence that describes the meaning of a term, and does not include the term itself (for an extended characterization of terminological definitions, see chapters 1 and 2 in Seppälä (2012)). The following examples illustrate terminological definitions from the domains of *chemical weapons* (Aureille 2003) and *oil spill cleanup* (Graf 2003).

5 Since terms are lexical units, their phonetic component is here implicit.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

| TERMS | DEFINITION |
|---|---|
| **distilled mustard**<br>**bis (2-chloroethyl) sulfide**<br>**sulfur mustard** | *A blister agent which is an amber brown liquid with an odor similar to that of burning garlic.* |
| **submunition**<br>**bomblet** | *A chemical munition of small size contained in a main one and designed to disperse non-persistent agents.* |
| **net boom** | *A boom that is made of netting to facilitate the retention of viscous oils.* |
| **pump system** | *The part of the skimmer that transfers the recovered oil, oil and water and/or emulsions from the skimmer head to a storage tank.* |

### 2.2 The Definition Writing Activity

For a definition-writing tool to be helpful, it should address the content-related questions that definition authors face during the definition writing process, such as how to select the relevant kind of information for defining a term (see section 2.4). In this section, I describe the different parts of the definition-writing activity (corpus compilation, information extraction, and definition construction) to introduce the main questions to which it gives rise.

When engaging in definition writing, terminologists generally gather a corpus of texts written by experts of the domain of interest (Example 1). This corpus provides the terms to be defined and the necessary evidence as concerns the terms' meanings, referents, and uses. They then identify and extract all the textual information units expressing features of the concept in the compiled corpus (text in bold in Example 1).

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

**5.3.2 Net booms**

**Netting** can be used as **a boom to collect viscous oils at sea**. Vessels deploy the netting, manoeuvring it to surround and concentrate the oil for recovery. In inshore waters, it may be possible to use nets as a moored boom. Net booming is based on the principle that **air and water will pass through the net but not viscous oil**. Therefore, there is little or no downward current to carry oil underneath the net and the net is less prone to being blown flat. **Because of the lower resistance of nets to water movement, it should also be possible in theory to deploy net booms in faster currents** than is possible with conventional booms. **Net booms have not yet been fully tested in an actual oil spill but results from field trials have been encouraging**.

A net boom **consists of a long strip of netting, supported at frequent and regular intervals by poles with floats and weights attached to keep the netting upright** (Figure 20).

Example 1: Sample of a corpus for defining the term *net boom* and all the information (in bold) related to that term

After that, they select those units that are the most relevant for defining the term as it is used in the domain under consideration. The pieces of information found in a specialized corpus can be sorted intuitively into three categories according to their relevance to the domain (Seppälä 2009):

- *Latent features* ($F_L$) are those parts of the expert's concept of the referent that are excluded from the dictionary entry or the ontology because they are irrelevant to the domain of expertise. For example, "Floating Net Booms are used for a wide variety of applications across various industries."[6] This piece of information may be in an expert's mind when considering net booms, but it is not relevant to be included in a specialized dictionary or an ontology of the oil spill cleanup domain.
- *Salient features* ($F_S$) are those parts of the expert's domain-specific concept of the referent which can be included in a dictionary entry or an ontology as they are of interest relative to the domain of expertise. For example, "Because of the lower resistance of nets to water movement, it should also be possible in

[6] Source of the citations: Elastec/American Marine, http://www.elastec.com/customcontainment/netboom (May 29, 2014).

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

theory to deploy net booms in faster currents" and "Net booms have not yet been fully tested in an actual oil spill but results from field trials have been encouraging." These pieces of information correspond to the sorts of features that are likely to be of interest to experts in the oil spill cleanup domain and may thus be included in a specialized dictionary or an ontology of this domain. However, they are not relevant for defining a net boom; rather, they would be included in a *note* or *comment* field.

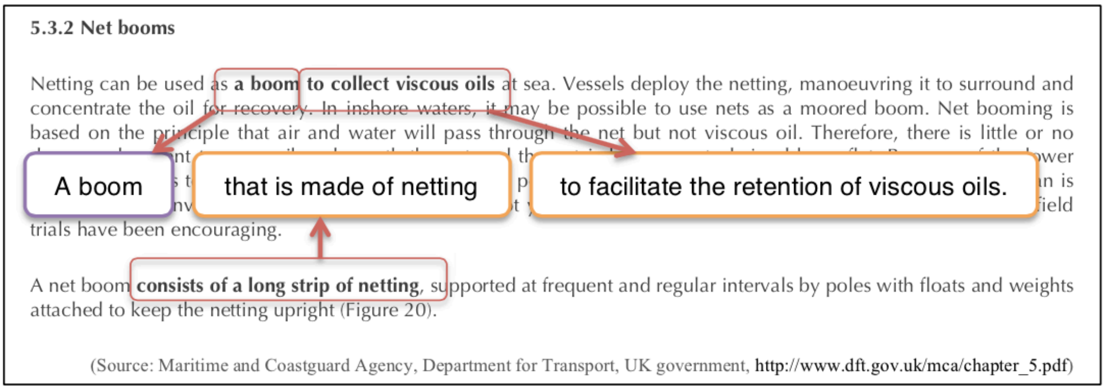
- *Potentially relevant features* ($F_{PR}$) correspond to a subset of salient features that are of interest relative to the domain of expertise, and that can be considered relevant for defining a term. For example: "Netting"; "a boom to collect viscous oils at sea."; "air and water will pass through the net but not viscous oil."; "consists of a long strip of netting, supported at frequent and regular intervals by poles with floats and wights attached to keep the netting upright."

Following this intuitive feature selection process, terminologists discard the *latent features* ($F_L$) since they are considered irrelevant to the dictionary entry. Then, from the set of *salient features* ($F_S$), they select the *potentially relevant* ($F_{PR}$) ones as candidates for composing a definition. Finally, they select a subset of the *potentially relevant* features as *relevant* ($F_R$) to be included in the definition, such as "a boom to collect viscous oils" and "consists of a long strip of netting" in Example 2. Therefore, a definition should express *relevant features* ($F_R$) among the *potentially relevant features* ($F_{PR}$) of the domain-specific concept of the referent.

In the last step, terminologists construct, with the selected units of information, a concise and informative sentence called a *definition* (Example 2). This definition is intended to help the dictionary users to become (more) competent in their

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

usage of the domain’s terminology, and data curators (more) competent in data annotation.




Example 2: Constructing a definition of the term *net boom* with the relevant features

## 2.3 The Internal Structure of Terminological Definitions

The resulting definition exhibits a conceptual structure composed of the features selected as relevant for defining the relevant term. Its underlying logical form is generally that of the classical Aristotelian definition of a *species* via *genus* and *differentiae* (as in: *a human* (species) *is an animal* (genus) *that is rational* (differentia)). Thus, a definition can be decomposed into two parts containing the corresponding textual information units, and fulfilling the genus and differentia roles respectively.

In the simplest case, the role of the genus is to state the type of thing to which the term refers by way of the implicit *is_a* relation, as in *a submunition* *is a chemical munition*, and *a net boom* *is a boom*. This is the strict Aristotelian genus case. But the part of the definition that plays the genus role can also express some other type of relation. It often expresses a parthood relation, as in *a pump system is the* *part of the skimmer* *that* …. It may also express other relations, as in *distilled mustard* *is a blister*

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

agent, where *blister agent* expresses the role of a chemical compound that is relevant in the domain of *chemical weapons*.

The role of the differentia is to distinguish the meaning of the defined term from that of its neighboring terms in the domain. It does so by expressing some feature of the referent of the defined term that is relevant to the experts of the domain. For example, the above definition of *net boom* contains two differentiae: *made of netting*, which expresses a physical property of the referent, and *facilitates the retention of viscous oil*, which expresses a function.

## 2.4 Questions Raised by Definition Writing

One of the most challenging tasks of definition writing is selecting the defining information. Indeed, when writing definitions, terminologists constantly face the following questions: *On what grounds is this selection made?* and *What kind of information is relevant in this definition?* To better understand the feature selection issue involved in definition writing, I specify a set of three more specific questions that have to be answered to develop definition-writing tools:[7]

**Q1** How to select the features that could potentially be considered as relevant in a definition? The first question addresses the problem of discriminating between what we can think of as *defining* (potentially relevant, $F_{PR}$) and *non-defining* (latent and salient, $F_L$+$F_S$) features.

**Q2** Among the set of *potentially defining features* ($F_{PR}$), how to select the *relevant features* ($F_R$), i.e., those that will be included in a particular definition?

[7] For a more technical account of these questions, and a more detailed description of why they are relevant and how they arise in the process of definition writing, see Seppälä (2009, 2010, 2012, chapter 3).

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

**Q3** How to select the appropriate genus of a definition when multiple candidates have been extracted from the corpus? For example, how to choose between *Netting* and *a boom* in our corpus example above.

To create a program that would help solve these selection problems, we thus have to:

- Identify clear and operational content selection principles that help definition authors (and, in consequence, the system) to identify the relevant kind of definition content.
- Find the best way to represent (model) definition contents so as to produce templates or rules that are understandable for definition authors and that can be implemented in a computer system.

These requirements will serve to assess the existing solutions to feature selection.

## 3 Available Means to Address Feature Selection Issues

We can distinguish two approaches to understand feature selection in definition writing: *quantitative* and *qualitative* approaches. We will see whether they can be used to provide clear and operational feature selection principles as a basis for creating generic definition templates.

Terminology traditionally sees itself as relying on the classical Aristotelian principles of necessary and sufficient conditions, or on the Roschian psychological notion of prototypicality.[8] I call both of these approaches *quantitative* because they

[8] In terminology, the principle of definition by necessary and sufficient conditions was historically advocated by Wüster and, more recently, by the Vienna School, and the ISO. Prototypicality has influenced different scholars (Temmerman 2000a; Zawada and Swanepoel 1994; Weissenhofer 1995), as well as different terminological approaches, like those inspired by *cognitive semantics* (Zawada and Swanepoel 1994), and those that combine it with *frame semantics*, such as the *sociocognitive approach* (Temmerman 2000b) or the *frame-based terminology* (Faber Benítez 2009). Weissenhofer (1995) even proposes integrating prototypicality into Wüster's four-field model.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

primarily account for the relationship between the intension and the extension of definitions, that is, between the contents of a definition and the set of individual referents encompassed by the definition. This relationship may be expressed in terms of a percentage of the members of an extension satisfying the intension: according to the classical Aristotelian approach, the coverage rate is 100%, as in *all triangles have three sides*; according to prototypical views, it is somewhere between >50% and <100%, as in *most swans are white*.

Neither of these principles provides implementable feature selection constraints for developing a semi-automated definition writing system. The Aristotelian definition principles cannot be successfully applied if one has no previous knowledge of how terms in a given domain differ from one another in meaning — as is often the case in professional settings where terminologists write definitions on a case-by-case basis applying their expertise in multiple domains. Distinguishing the defined term from its neighboring terms by means of necessary and sufficient conditions requires some acquaintance with the overall organization of the conceptual system to which the definition belongs (see, for example, ISO 704 2009). The prototype approach relies on a similarity principle that tells us to select the most typical features with respect to the prototype of a category. While this would answer question Q2 on the selection of potentially relevant features among salient features, it does not provide any answer to the other two questions. It can therefore not be used for formalizing the selection of defining features. Both approaches mainly account for the relationships between the intension and the extension of the definition; neither of

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

them meets the requirements for operational content selection principles and for definition modeling by way of predictive templates.[9]

Concerning content-related matters, many authors have adopted a relational perspective on the features composing terminological definitions:[10] the contents of the definitions are described in terms of relations between the differentia and the genus or the defined term, as schematized in Figure 1.

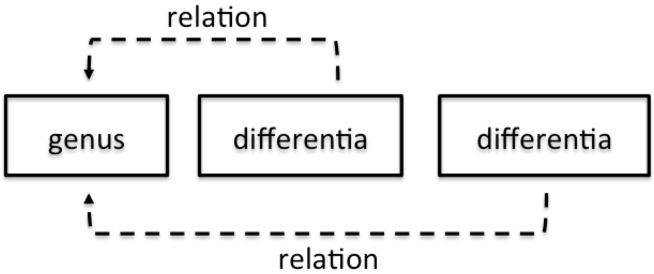



Figure 1: Relational description of definition contents: each differentia expresses some relationship to the genus or to the defined term

The first formalizations and methodological frameworks for analyzing terminological definitions in relational terms were proposed by Sager and colleagues (Sager and L'Homme 1994; Sager and Ndi-Kimbi 1995).

The idea behind these *qualitative* approaches is to analyze regularities in the internal structure of definitions in terms of relations (for example, *has_part*, *has_function*, etc.), and to provide explanations for why some types of information are (to be) included in a definition.[11] Thus, these approaches use some kind of relational labeling to formalize the contents of the definitions — lexico-semantic,

---

[9] For a discussion of other limits of these quantitative approaches for modeling definition contents, see Seppälä (2012, chapter 4).

[10] See, for example, Dancette (2005), Feliu (2000, 2001), Feliu and Cabré (2002), Feliu (2004), L'Homme (2007), Sager and Kageura (1994), Sager and L'Homme (1994), Sager and Ndi-Kimbi (1995), García de Quesada (2001), García de Quesada et al. (2002).

[11] For lack of space, it is not possible here to present the relational approaches that mostly contribute to the advances in the formalization of definitions in terminology. For a review, see chapter 4 in Seppälä (2012).

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

conceptual, and ontological relations. In the lexico-semantic approaches, the definition contents are formalized, for example, by means of *actantial roles* (*agent*, *patient*, etc.) and *lexical functions* (for example, $S_{loc}$ for *typical location*) designed to explicate the defined term's relations to other terms (Dancette 2005; L'Homme 2007). As lexico-semantic relations tend to be language-dependent, they are not suited for the current definition content modeling endeavor. In the conceptual approaches, the contents of definitions are mostly represented in the form of *relational definition models* or *templates* that specify the types of information (expressed as relations) that are relevant for defining terms belonging to a given category, such as OBJECT, PROCESS, etc.[12] In most of these works, the categories and relations used for formalizing the definition contents are partly determined via top-down methodologies using a restricted number of upper-level categories that are complemented via bottom-up methodologies such as domain-specific corpus analyses (Faber 2002; Faber Benítez 2009). However, the relational definition models resulting from the latter approaches are also unsuited for our purposes: they are based on domain-specific conceptual categories that can be created in an *ad hoc* manner (Smith 2004). As mentioned earlier, the objectives of this research are bound by a *genericity constraint*: what we are seeking is a generic tool to assist terminologists and ontologists in definition writing, and this tool should be domain- and language-independent. The corpus-based bottom-up methodologies should nevertheless be useful to determine the corresponding domain-dependent categories and relations, and to constrain the generic definition content models to yield more relevant models with respect to a particular domain or context (see section 6.2).

[12] Feliu (2000, 2001), Feliu and Cabré (2002), Feliu (2004), Sager and Kageura (1994), Sager and L'Homme (1994), Sager and Ndi-Kimbi (1995), García de Quesada (2001), García de Quesada et al. (2002).

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

Regarding the factors constraining feature selection in definitions, we can identify four main explanatory approaches that have contributed to the advancement of definition modeling and to the principles guiding feature selection in the terminological literature:

- Conceptual and other approaches deriving from ontological research: selection of relevant features depends partly on the type of category of the defined term and on the context (domain's conceptual system or communicative setting) in which the definition is produced (García de Quesada et al. 2002, Sager and L'Homme 1994).
- Lexico-semantic approaches drawing on, e.g., Mel'čuk's theory of *Explanatory Combinatorial Lexicology*: selection of relevant features depends on the linguistic system of the language to which the defined term belongs (Dancette 2005; L'Homme 2007).
- Cognitive approaches drawing on Langacker's *Cognitive Semantics*, on Fillmore's *Frame Semantics*, and on Lakoff's *Idealized Cognitive Models*: selection of relevant features depends on "the actions and processes that take place in the specialized field" (Faber Benítez 2009, 123-124; Faber Benítez 2009; Temmerman 2000b).
- A pragmatic approach: selection of relevant features depends not only on the type of category of the referent and the domain's conceptual system, but also on the users' specific needs (García de Quesada 2001).

The shortcoming of these approaches is that they tend to focus on a few, if not only one, of the possible explanations to feature selection; even the pragmatic

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

approach proposes methodological solutions that are not generic.[13] Despite these limitations, the relational approaches have the advantage of dealing both with definition content modeling and feature selection principles; they meet the requirements (specified in section 2.4) for implementing definition-writing aids. This work thus builds up on existing relational approaches. The main differences lie in the adoption of a top-down methodology to determine the categories and relations constituting the generic relational models (section 6.1), and in the integration of all the acknowledged feature selection factors together into a single framework (section 4).

## 4 A Proposal Toward an Integrated Approach to Feature Selection

The review of the different relational approaches to definitions in terminology shows that, while feature selection can be explained by distinct factors (conceptual/ontological; systemic; communicative; linguistic), these are nonetheless complementary: these factors determine the relevance of the defining features in various degrees. I therefore propose a comprehensive and structured account of these factors along three dimensions, each of which includes at least two constraining factors:

1 *Extensional dimension*:

  a *Ontological factor*: feature selection is influenced by the type of referent of the defined term (see section 5 below).

  b *Type vs. instance factor*: feature selection is influenced by whether the referent is a type of thing or a particular instance of such a type.

[13] García de Quesada (2001) does acknowledge the influence of all of these explanatory factors, but focuses on the development of relational models for definitions where the default constraints on feature selection are conceptual, thus domain-dependent, and can be overridden by the end-user.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

c *Extensional factor*: feature selection is influenced by whether the extension is composed of homogeneous or heterogeneous things.

2 *Contextual dimension*:

a *Systemic factor*: feature selection is influenced by the conceptual system.

b *Individual factor*: feature selection is influenced by the background knowledge of the dictionary/ontology user.

3 *Communicative dimension*:

a *Functional factor*: feature selection is influenced by the specific needs of the dictionary/ontology user for example when engaging in a specific task, such as normalization which requires restricting the meaning of terms to exactly what is in the definition to avoid ambiguities.

b *Situational factor*: feature selection is influenced by the communication situation (for example expert vs. non-expert).

Each factor yields hypotheses that need to be tested in order to gauge their influence on feature selection. However, each hypothesis requires distinct methodologies for testing, and a common descriptive framework so that the influence of the respective factors can be compared. The existing methodological approaches cannot currently meet the latter requirement since they yield domain-dependent descriptive frameworks. I will present in section 6 an *ontological analysis framework* in order to address this problem. This framework will provide a language- and domain-independent descriptive vocabulary that different methodological approaches can use to test the respective influence of the various acknowledged feature selection constraints. Carrying out these different tests in a systematic and integrated fashion should yield empirical results that would give a better understanding of the definition writing and feature selection processes. Taking into account these different

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

dimensions simultaneously would at the same time provide the basis for an integrated theory of definitions in terminology.

The next section specifies the research hypotheses relating to the constraining factors that should prove the most fruitful for the creation of language- and domain-independent defining models, that is, the hypotheses related to the ontological factor (1a).

## 5 Research Hypotheses

The main hypothesis (H0) of this research requires taking into account all of the factors known or assumed to influence feature selection, and ruling out those which have a low likelihood of allowing for the development of generic templates. The hypothesis reads as follows:

> **H0** *While feature selection is constrained by different factors, only the ontological factor constitutes a basis to the formalization of feature selection in definitions that is independent of the domain and the language of the definition.*

This hypothesis is supported by *a priori* considerations. According to the theoretical background, conceptualizations may vary according to domains and, possibly, languages or socio-linguistic contexts. This rules out the factors under the contextual and communicative dimensions. The *type vs. instance* and the *extensional factors* under the extensional dimension could be considered, but they are rather complementary factors to consider, as they do not lend themselves to a relational modeling of definition contents.

Therefore, considering all the factors that may constrain feature selection, the one that should, *prima facie*, be the most *neutral* regarding domains and languages is the *ontological factor* (1a) according to which feature selection partially depends on

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

the type of entity to which the defined term refers. The other factors can be considered in a further phase, in order to create, for example, domain-adaptable tools. I thus concentrated my research on this ontological factor, and on the corresponding hypothesis H0.

Having identified the one factor that satisfies the above-mentioned *genericity constraint*, we can ask, in view of our goals, two more specific research questions: (1) *To what extent is it possible to establish domain- and language-independent (generic) definition models based on the type of entity that is the defined term's referent?*; (2) *Is it possible to constrain these generic models to predict the most relevant features for defining a term in a specific context?*

To answer these questions, I make the following testable hypotheses:

**H1** *Feature selection is partially determined by the entity type to which the defined term refers.*

**H2** *Generic relational models can be constrained to obtain relevant models for each entity type.*

In the following section, I present an ontological analysis framework and a specific methodology to test these hypotheses.

## 6 Creating Generic Models and Increasing Their Relevance

The methodology for testing hypotheses H1 and H2 involves empirical corpus studies of existing terminological definitions annotated within the proposed ontological analysis framework. This framework involves creating domain- and language-independent generic relational models based on ontological categories.

By 'ontology' we mean: "a representation of the categories of objects and of the relationships within and amongst categories that are to be found in any domain of reality whatsoever." (Spear 2006, 28) The sort of formal, domain-neutral, upper-level

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

ontology represented by BFO, SUMO, DOLCE, etc. Here, the ontological categories are those of the realist upper-level ontology *Basic Formal Ontology* (BFO, 2014) (see section 6.1.1). They thus differ from the domain-specific conceptual categories used in similar terminological works. The relational models are aimed at formalizing and predicting feature selection, therefore definition contents, whatever the terminological context. The vocabulary in the models constitutes a domain- and language-independent metalanguage used for describing the internal structure of definitions. The corpus studies are intended to test the predictability of the models and to establish whether they can be constrained in such a way as to be relevant to a specific definition task/domain.

The corpus studies consist in the analysis of definitions extracted from multiple domains and written in different languages. The internal (relational) structure of definitions is analyzed by means of the BFO categories and their relations to other BFO categories. The underlying assumption of this methodology is that the relational structure of existing definitions is a reliable source for establishing whether the type of entity to which the defined term refers has an influence on the type of content expressed in specialized definitions.

### 6.1 Creating Generic Relational Models

Traditional conceptual analysis approaches to studying definition contents in terminology require defining a set of *categories* and *relations* that characterize these categories (Sager and Kageura 1994; Sager and L'Homme 1994). The novelty of the proposed ontological approach is that the categories (entity types) and their corresponding relations are not conceptual, nor are they established or completed through corpus analysis, as it is generally the case in terminology (Faber 2002; Faber Benítez 2009; García de Quesada 2001, García de Quesada et al. 2002; Kageura

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

2002). Instead, this framework uses a set of categories corresponding to the types of entities that exist in the world, to which defined terms refer. Each entity type is characterized by a *relational model*, i.e., a set of *relational configurations* (RCs) expressing its relations to other entity types (see section 6.1.2).

#### 6.1.1 Using the BFO Categories to Model Definition Contents

The rationale behind the choice of BFO is that this ontology is aimed at representing the types of things that exist in the world, their properties, and their relations to other types of entities, regardless of any domain-specific conceptualization. Entity types are organized according to philosophical distinctions (Smith 2003; 2012), and they are consistent with scientific knowledge of the world. BFO is indeed aimed at defining the elementary building blocks for the ontological representation of domain-specific ontologies, such as are used in bio-medicine and related disciplines.[15] BFO is currently used by over 100 ontologies which are thus interoperable.[16]

The choice of BFO is in line with the assumed theoretical background: definitions describe types of things or, sometimes, particular things in the world in ways which seek to capture how experts in different domains conceptualize them, and for which experts have specific terms. As a realist upper-level ontology, BFO is also adapted for testing hypothesis H1 according to which the selection of relevant features to define a term is partly influenced by the type of entity to which the defined term refers. Moreover, analyzing definitions within this framework makes it possible to see to what extent domain-experts judge other types of contents relevant for defining terms; this may reveal typical features for defining different categories. In this respect,

[15] BFO does not include entities of any specific domain. It is intended as an upper‑level ontology to describe any type of specific domain, thus rendering domain-specific ontologies compatible and interoperable (Smith and Ceusters 2010).

[16] For a list of ontologies and projects using BFO, see http://www.ifomis.org/bfo/users (May 29, 2014).

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

the BFO categories and relations constitute a basic fixed and well-defined, descriptive, language- and domain-independent controlled vocabulary (or metalanguage) that fills a gap often deplored in terminology (Sager 1990, 28-29).

The first version of the BFO-Definition-Models proposed in this work is based on the thirty-five categories of BFO 1.0, the first version of BFO, and its specifications (Spear 2006). This version lacked an entity type that is a common referent of terms in domains like banking, commerce, and administrative systems. I therefore completed the set of BFO-based models with a model for QUASI-ABSTRACT ENTITIES based on articles by Barry Smith and his colleagues (Smith 2008, 2011, 2014; Smith and Ceusters 2010). Also, as BFO cannot deal well with fictions, and with mathematical and other abstract entities, I created a tentative category for ABSTRACT ENTITIES.[17] With these additions, the total number of relational models adds up to thirty-seven.

#### 6.1.2 Contents of the Relational Models

Each relational model consists of the relational configurations (RCs) that characterize the corresponding category (Example 3). Each RC has two parts:[18]

(1) A unidirectional relation (*in_lowercase_italics*) between the category of the model and a second category called the relatum, and

(2) The relatum itself (IN SMALL CAPS).

Since the categories are hierarchically organized, the RCs at the lower levels inherit the RCs from their parent levels. Example 3 shows the relational model for the category OBJECT. Two RCs specifically characterize this category; the five other RCs

---

[17] The BFO community is working on the *Information Artifact Ontology* (IAO) to address these matters.

[18] Relational configurations (RCs) can also be considered as triples composed of ENTITY TYPE+*relation*+ENTITY TYPE. To avoid redundancy on the left part of the models, I describe here the RCs as pairs composed of *relation*+ENTITY TYPE.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

are inherited from the category INDEPENDENT CONTINUANT, which is the direct parent of OBJECT in BFO 1.0. Thus, according to this model, 'every OBJECT *has_part* some OBJECT', and 'every OBJECT *bearer_of* some QUALITY'.

| **OBJECT** |
| --- |
| **RCs characterizing the entity type OBJECT** |
| *has_part* OBJECT<br>*participates_in* PROCESS |
| **RCs inherited from the entity type INDEPENDENT CONTINUANT** |
| *bearer_of* QUALITY<br>*bearer_of* REALIZABLE ENTITY<br>*located_at* TEMPORAL REGION<br>*located_in* SITE<br>*participates_in* PROCESSUAL ENTITY |

Example 3: Relational model for the entity type OBJECT

### 6.2 Making the Generic Models Relevant

To test whether the generic models are predictive and whether it is possible to make them relevant, the method involves analyzing a corpus of existing terminological definitions. Each definition is segmented according to the general *genus*+*differentiae* pattern; then, each segment is tagged with the corresponding predefined category for the genus, and relational configuration for each differentia. In a second step, the annotated structure of the definitions is compared to each type's relational model to see if its RCs are expressed in the definitions and which RCs are the most frequent.

The examples below illustrate terminological definitions annotated according to this ontological annotation schema. REF indicates the real referent of the defined term when it is not already expressed by the genus (for example, a chemical compound defined as a doping agent in sports); GEN indicates the segment having the *genus* role in the definition; SPE indicates a segment having a *differentia* role; *other* indicates that the RC expressed by the segment is not part of the category's relational model. In the latter cases, a new RC is proposed using the BFO categories and

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

relations as a controlled vocabulary. All the RCs in the models are unidirectional; they are to be understood as a relation holding between:

- the referent of the defined term and the ontological category to which the annotated segment refers, for GEN and SPE constituents (Examples 5 and 6);
- the referent of the defined term and the ontological category to which it pertains, when the GEN has a different referent (see REF in Examples 4 and 7).

| **distilled mustard**<br>**sulfur mustard**<br>**bis (2-chloroethyl) sulfide** | | *A blister agent which is an amber brown liquid with an odor similar to that of burning garlic.* |
|---|---|---|
| REF | *is_a* OBJECT | |
| GEN | *bearer_of* REALIZABLE ENTITY | A blister agent |
| SPE | *bearer_of* QUALITY | which is an amber brown liquid with an odor similar to that of burning garlic. |

Example 4: Annotation of the internal structure of the definition of *distilled mustard* in the domain of *chemical weapons* (Aureille 2003)

| **submunition**<br>**bomblet** | | *A chemical munition of small size contained in a main one and designed to disperse non-persistent agents.* |
|---|---|---|
| GEN | *is_a* OBJECT | A chemical munition |
| SPE | *bearer_of* QUALITY | of small size |
| SPE | *located_in* SITE | contained in a main one |
| SPE | *bearer_of* REALIZABLE ENTITY | and designed to disperse non-persistent agents. |

Example 5: Annotation of the internal structure of the definition of *submunition* in the domain of *chemical weapons* (Aureille 2003)

| **net boom** | | *A boom that is made of netting to facilitate the retention of viscous oils.* |
|---|---|---|
| GEN | *is_a* OBJECT | A boom |
| SPE | *has_part* OBJECT | that is made of netting |
| SPE | *bearer_of* REALIZABLE ENTITY | to facilitate the retention of viscous oils. |

Example 6: Annotation of the internal structure of the definition of *net boom* in the domain of *oil spill cleanup* (Graf 2003)

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

| **pump system** | *The part of the skimmer that transfers the recovered oil, oil and water and/or emulsions from the skimmer head to a storage tank.* |
|---|---|
| REF | *is_a* OBJECT | |
| GEN | *other*: *part_of* OBJECT | The part of the skimmer |
| SPE | *bearer_of* REALIZABLE ENTITY | that transfers the recovered oil, oil and water and/or emulsions from the skimmer head to a storage tank. |

Example 7: Annotation of the internal structure of the definition of *pump system* in the domain of *oil spill cleanup* (Graf 2003)

## 7 Case Study: A Preliminary Corpus Analysis

To evaluate the applicability of the proposed ontological analysis framework and its methodology, and get preliminary results, I conducted a pilot corpus study (Seppälä 2012).

### 7.1 A Corpus of Terminological Definitions

The corpus consists of a sample of 240 terminological definitions randomly extracted from a larger corpus comprising 1495 definitions written in French. The definitions were extracted from multilingual terminological dictionaries that cover 15 domains in a systematic way. Graduate-level students in terminology wrote these dictionaries as their diploma work. The students had an extensive training in definition writing, and did a mandatory internship in terminology during which they wrote definitions to be included in the databases of their employers.[19]

The quality of the definitions was both indirectly and directly controlled: indirectly, through the selection of works that were allotted the minimum grade required to obtain the diploma in terminology (4/6, 6 being the highest score); directly, by automatically checking their compliance with the generally accepted

[19] For a justification of the use of these dictionaries, see Seppälä (2012, section 6.3). For a discussion on the use of students' terminology works for other purposes, see Candel and Humbley (2002).

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

formal rules of definitions (Seppälä 2006a, 2006b). Sixty-eight definitions that did not comply with these rules were thus excluded.

From the 37 relational models, 16 categories and 73 RC types (462 RC occurrences) are represented in the analyzed corpus. This reduced number of categories can partly be explained by the absence of terms referring to entities belonging to 10 models that represent absolute space and space-time: SPATIAL REGION (5 models) and SPATIOTEMPORAL REGION (5 models). These entity types are considered as containers "within which substances and their qualities exist" or "within which processes unfold" (Spear 2006, 60). They appear in domains like physics, which are not covered by the analyzed corpus.

## 7.2 Corpus Analysis

The corpus analysis consists in segmenting and annotating each definition according to the annotation schema presented in section 6.2. The segmentation of the genus was automated and manually checked during the annotation. The details of the automatic genus tagging process are presented in Seppälä (2007). The annotation includes tags indicating whether the categorization of the segment or of the referent of the defined term is uncertain (see column 3 in Table 1).

Once all the definitions were annotated, I carried out frequency analyses to determine which relational configurations are used to define the categories in the corpus. The results for each category were then compared to the corresponding relational models to see which RCs are relevant for defining each category.

Table 1: Results for the entity type OBJECT

| **OBJECT** | **normal** | **categ.** | **other** | **total** | **%** |
|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) |
| ***has_part*** **OBJECT** | 15 | 1 | | 16 | 26 |
| ***participates_in*** **PROCESS** | 3 | 6 | | 9 | |

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

| | | | | | |
|---|---|---|---|---|---|
| ***bearer_of*** **QUALITY** | 19 | 3 | | 22 | 60 |
| ***bearer_of*** **REALIZABLE ENTITY** | 31 | 1 | | 32 | |
| *located_at* TEMPORAL REGION | 0 | 0 | | 0 | |
| *located_in* SITE | 2 | 2 | | 4 | |
| *Other* | | | 14 | 14 | 14 |
| **total** | **70** | **13** | **14** | **97** | **100** |

Table 1 is a result example for the definitions of terms referring to the category OBJECT. It shows that the generic model for OBJECT is highly predictive: 86% of the segments are described with one of the RCs of the model (column 6). This result is reliable since in most of the cases (70/97 segments) the annotation with the RCs didn't raise any particular categorization problem (column 2). Column 5 also shows which RCs (**in bold** in column 1) are the most relevant. Thus, OBJECTS are preferably defined in terms of their functions, roles, and qualities (*bearer_of* REALIZABLE ENTITY and *bearer_of* QUALITY), as well as in mereological terms (*has_part* OBJECT); whereas physical or temporal locations seem to be rather irrelevant.

### 7.3 Results

This preliminary corpus analysis yielded very promising results: almost 75% of the relational configurations expressed in the analyzed definitions belong to the models associated with the represented entity types (hypothesis H1). This shows that the generic models tend to be predictive independently of the domain.[20] Moreover, this empirical study shows which RCs in these generic models are most relevant in specialized definitions (hypothesis H2). Therefore, the pilot study suggests that it is possible (i) to create the needed generic models by using realist upper-level

[20] It might still be the case that some domains cover more of one type of entity than of others.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994

ontological categories to predict definition contents, and (ii) to make these models relevant by analyzing existing definitions.

The results also show in how many and in which cases the actual referent of the defined term is not referred to explicitly in the definition: around 25% of the analyzed definitions use a genus (GEN) that expresses another relationship than *is_a* to the relatum, or have no genus at all — e.g., in the case of definitions of QUALITIES and REALIZABLE ENTITIES generally expressed by adjectives. These cases reflect the fact that the organization of conceptual systems can differ from a strictly realist ontological organization. This finding might be an incentive to carry out comparable studies of the systemic feature selection factor (2a) using the same descriptive vocabulary to see when and why that is the case.

The results can further prove useful for assigning weights to the relational configurations of a model, thus indicating their relevance level, and for revealing *typical* RCs used for defining different entity types. The latter entails using BFO categories and relations as a controlled descriptive vocabulary (or metalanguage) to describe the contents of the segments marked *other*.

### 7.4 Limitations of the Pilot Study

This corpus analysis being a pilot study, it has some limitations. Above all, the corpus is only composed of definitions written in French. To test for the language-independence of the models, these analyses should also be applied to definitions written in other languages. The data gathered to extract the pilot corpus do however compose a large multilingual corpus of definitions of equivalent terms written in different languages — most of the time, in two. The larger multilingual corpus can be used to test for this linguistic constraint. In the meantime, however, the pilot corpus can be used as training data for (semi-)automatically annotating larger corpora.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

Finally, this empirical work also raised a number of questions that I discuss in the next section.

## 8 Discussion and Future Work

The proposed methodology raises some difficulties regarding the creation of the models and their applicability to the description of the internal structure of specialized definitions. I discuss them in relation with the models based on BFO and the above pilot study. I also discuss possible worries concerning the underspecification of the models and their usefulness for feature selection.[21]

### 8.1 Application Difficulties

First, the realist (upper-level) ontology used for creating the models might not be exhaustive. That was the case in the work described here. The entities that correspond to the ABSTRACT and QUASI-ABSTRACT ENTITY models are absent from BFO 1.0 and its specifications used for the pilot study. The same applies to relations, i.e., the left italicized part of the relational configurations: some of these are mentioned in the specifications, but not defined; others are defined by the authors of BFO elsewhere (Smith et al. 2005; Smith and Grenon 2004). Finally, some relations such as *agentive participation* and *causality* that would seem intuitive (at least from the conceptual and lexico-semantic analysis perspectives) are also absent from BFO. However, they can easily be added to the proposed ontological analysis framework by using for example participation-type relations from the *Relation Ontology* (RO) that extends BFO.

These difficulties are discussed with BFO experts in light of the new version, BFO 2.0. A further consolidation process also involves updating and enhancing the models with the specifications of BFO 2.0, where most of the terminological issues

[21] Other difficulties, not discussed here, might arise if the methodology is used with another realist upper-level ontology, although there is no such candidate yet, or with non-realist ontologies.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

have been solved. The new version contains some modifications in the hierarchy and terminology, and introduces relations as well as new entity types that should account for the QUASI-ABSTRACT ENTITIES introduced in this pilot study. The models are being updated and tested.

Second, the annotation of terminological definitions sometimes requires domain expertise the lack of which may yield categorization uncertainties (here marked *categ.*). Categorization difficulties may also be related to somewhat technical or too brief specifications of the categories and relations in the models; to the non-exhaustivity of the ontology used for creating the models; and to limited diversification of examples.

As with any conceptual and semantic annotation framework, to minimize categorization difficulties, annotators should be provided with a comprehensive annotation manual, including definitions, alternative terminology, and diversified examples to guide them in the proper use of the models. Current work includes the creation of such an annotation manual for the relational models based on BFO.

### 8.2 Underspecification of the Models for Feature Prediction

A more general concern would be that, contrary to other modeling approaches, the proposed models might be underspecified, and would thus fail to predict the distinctive features needed for definitions within a particular conceptual system. This is a legitimate worry, since the models are meant to represent upper-level domain-independent entity types and their characteristics; not domain-specific ones. One may thus object that the relational configurations in the models are so general — hence the underspecification of the models — that they are not useful. According to this objection, useful models should predict definition contents in a more fine-grained manner. They should account for the SPE segments (described here in terms of

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

broader RCs) that have a substructure. Indeed, this substructure might contain a sub-feature that is actually the one relevant for distinguishing terms of a particular domain. However, the results of the pilot study suggest that the BFO-based generic models are highly predictive, which means that they may often indirectly predict those more specific distinctive sub-features. Therefore, even if underspecified, generic models should still capture the features relevant to the domain.

Furthermore, for predicting defining features relevant in a specific domain, other methodologies should be used to test the other hypotheses on feature selection (see section 3). Systematically studying and comparing other factors would then yield a better understanding of the influence of, for example, conceptual systems on the contents of definitions. However, for those other tests to be relevant, they should be compatible with one another, and with the one presented here. Whenever more specific categories and relations are required to study domain-specific aspects, the descriptors should be based on the categories of BFO or otherwise linked to them, so that the metalanguage remains coherent. Domain-specific descriptive frameworks should thus be extensions of this upper-level ontology to allow for comparative studies.

### 8.3 Usefulness of the Models for Feature Selection

Another concern relates to the usefulness of the models for selecting the relevant information among all the pieces of information extracted from experts' texts. Indeed, these texts might only express the type of information (relational configurations) found in the relational models. However, this would have no incidence on the usefulness of the models in a (semi-)automatic tool to assist in definition writing. The results of the corpus analyses can still be used for assigning weights to generate different definitions, with different levels of relevance. Also, if research revealed that

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

experts' texts do only include RCs from the models, this would be an interesting empirical verification of the importance of the referents in specialized conceptualizations. It would show that meaning is to a large extent influenced by reality and its structure — and not only by the internal organization of conceptual or semantic systems. This, in turn, would indicate that feature selection should be based on, or completed with, other selection criteria from the extensional dimension, for example, to discard all information about particular things — as opposed to types of things. This would be a supplementary motivation to test the other hypotheses on feature selection. Nevertheless, the pilot corpus analysis presented here shows that 22.9% of the relational configurations expressed in the definitions are not included in the models. Since these defining features have also been selected among a broader set of features extracted from experts' texts, it is reasonable to expect to find in these texts pieces of information that are not already in the models. In any case, as this is an empirical question, it has to be tested on real data. The ideal setting to test this would be to constitute a corpus of existing definitions and of the original texts that were used to write these definitions as in the corpus study presented in Seppälä (2010, 2012, chapter 3).

# 9 Conclusion

In this paper, I addressed the troublesome question of feature selection and content prediction in definition writing. I showed that it is possible (i) to create generic — language- and domain-independent — relational models for predicting definition contents whatever the context, and (ii) to make these models relevant for each category in particular contexts or domains.

The main hypothesis studied in order to attain generic models is that *feature selection is partially determined by the entity type to which the defined term refers*

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

(H1). The second hypothesis relates to the possibility of constraining these models so that they predict relevant information: *Generic relational models can be constrained to obtain relevant models for each entity type* (H2).

To test these hypotheses, I proposed an *ontological analysis framework* and a methodology involving two parts: elaborating the generic relational models by using the categories of a realist upper-level ontology, here the *Basic Formal Ontology* (BFO), and their characteristics; and annotating and analyzing a large corpus of multilingual and multi-domain definitions. The corpus analysis is intended to show (H1) to what extent the characteristics of the referents of the defined terms — the relational configurations (RCs) composing the corresponding models — influence de contents of the definitions, and (H2) which RCs of the models are deemed relevant for each entity type in existing definitions.

To assess the proposed methodology, I presented a pilot corpus study that also shows what kind of results can be expected concerning the tested hypotheses. These results suggest that it is possible to create *generic* yet *relevant* relational models that can, for example, be implemented in a (semi-)automatic tool to assist terminologists and ontologists in definition writing.

The BFO-vocabulary used in the presented models is intended to be used as a basic fixed and well-defined, descriptive, language- and domain-independent controlled vocabulary (or metalanguage). This vocabulary could be fruitfully used to test other hypotheses on feature selection so as to obtain comparable results, and thus determine to what extent definition contents are influenced by the different constraining factors. It is not intended as a replacement of existing conceptual and lexico-semantic descriptive frameworks. It is proposed as a complementary analysis

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

framework that should be linked to the latter ones in order to further our understanding of the objects of terminology.

Using this ontological analysis framework would, furthermore, allows for easier domain-ontology development through an enhanced integration of textual definitions. Indeed, terms can be more easily linked together by producing definitions with ontological metadata. This linking would produce domain-ontologies that are compatible with other ontologies using the same upper-level ontology. The proposed relational models can therefore be of interest to terminologists, ontologists, and anyone seeking a unified framework to describe and compare terms, concepts, and even theories.

## 10 Acknowledgements

Work on this paper was supported by the Swiss National Science Foundation (SNSF). Thanks are due also to Barry Smith, Neil Otte, Jonathan Bona, Patrick Ray, and the anonymous reviewers for their comments and suggestions on this paper.

## 11 References

Aureille, Marie-Cécile. 2003. *Terminologie du nettoyage des déversements d'hydrocarbures / Oil spill cleanup terminology: anglais-français, English-French.* Genève: École de traduction et d'interprétation, Université de Genève.

Aussenac-Gilles, Nathalie, Sylvie Despres, and Sylvie Szulman. 2008. "The TERMINAE Method and Platform for Ontology Engineering from Texts." In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, edited by Paul Buitelaar and Philipp Cimiano, 199-223. Amsterdam: IOS Press.

The *Basic Formal Ontology* (BFO), http://www.ifomis.org/bfo (May 29, 2014).

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

Blanchon, Élisabeth. 1997. “Point de vue sur la définition.” *Meta* 42 (1):168-173.

Cabré, Maria Teresa. 1998. *La terminologie: théorie, méthode et applications*. Paris, Ottawa: Armand Colin/Les Presses de l'Université d'Ottawa.

Candel, Danielle, and John Humbley. 2002. “Réutilisation des mémoires d'étudiants en terminologie.” *Cahiers de Lexicologie* 80:153-166.

Célestin, Tina, Gilles Godbout, and Pierrette Vachon-L’Heureux. 1988. *Méthodologie de la recherche terminologique ponctuelle : Essai de définition*. Québec: Gouvernement du Québec.

Clas, André, éd. 1985. *Guide de recherche en lexicographie et terminologie*. Paris: Agence de coopération culturelle et technique.

Dancette, Jeanne. 2005. “Les représentations lexico-sémantiques (RLS), moyen de structuration des connaissances dans les domaines spécialisés.” In *ISKO-France, Grenoble*: 83-96.

Dubuc, Robert. 1978. *Manuel pratique de terminologie*. Montréal, Paris: Linguatech, CILF.

Faber Benítez, Pamela. 2009. “The Cognitive Shift in Terminology and Specialized Translation.” *MonTI, Monografías de Traducción e Interpretación* (1):107-134.

Faber, Pamela. 2002. “Terminographic Definition and Concept Representation.” In *Training the Language Services Provider for the New Millennium*:ed. by Belinda Maia, Johann Haller, and Margherita Ulyrich, 343-354. Porto: Universidade do Porto.

Feliu, Judit. 2000. *Relacions conceptuals i variació funcional: elements per a un sistema de detecció automàtica*. Trabajo de investigación presentado en el

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

Feliu, Judit. 2001. "Propuesta de clases y relaciones conceptuales: recopilación y análisis." In *La terminología científico-técnica : reconocimiento, análisis y extracción de información formal y semántica (DGES PB96-0293)*, edited by Maria Teresa Cabré and Judit Feliu, 143-163. Barcelona: Institut universitari de lingüistica aplicada, Universitat Pompeu Fabra.

Feliu, Judit. 2004. *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Feliu, Judit, and Maria Teresa Cabré. 2002. "Conceptual Relations in Specialized Texts: New Typology and an Extraction System Proposal." In *TKE2002, Terminology and Knowledge Engineering Proceedings, 6th International Conference*, Nancy, 45-49.

García de Quesada, Mercedes. 2001. "Estructura definicional terminográfica en el subdominio de la oncología clínica." *Estudios de Lingüística Española (ELiES),* Volumen 14, http://elies.rediris.es/elies14/ (July 25, 2014).

García de Quesada, Mercedes, Pedro A. Fuertes Olivera, and Silvia Montero. 2002. "Propuesta de estructura definicional terminográfica en OntoTerm®." *Terminology* 8 (1):57-90.

The *Gene Ontology* (GO): http://www.geneontology.org (July 25, 2014).

*Le grand dictionnaire terminologique*: http://www.granddictionnaire.com (May 29, 2014).

Grabar, Natalia, Thierry Hamon, and Olivier Bodenreider. 2012. "Ontologies and Terminologies: Continuum or dichotomy?" *Applied Ontology* 7 (4):375-386.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

Graf, Julie. 2003. “Glossaire sur les armes chimiques = Glossary on chemical weapons : français-anglais, French-English.” Genêve: Université de Genève, École de traduction et d'interprétation.

The *Infectious Disease Ontology* (IDO): http://infectiousdiseaseontology.org/page/Main_Page (July 25, 2014).

ISO 704. 2009. *Terminology Work — Principles and Methods (ISO 704:2009)*. 3rd ed. Geneva: ISO.

Kageura, Kyo. 2002. *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth.* Amsterdam: John Benjamins.

L’Homme, Marie-Claude. 2007. “Using Explanatory and Combinatorial Lexicology to Describe Terms.” In *Selected Lexical and Grammatical Topics in the Meaning-Text Theory. In Honour of Igor Mel’čuk*, edited by Leo Wanner, 167–202. Amsterdam, Philadelphia: John Benjamins.

Larivière, Louise. 1996. “Comment formuler une définition terminologique.” *Meta* 41 (3):405-418.

Marconi, Diego. 1997. *Lexical Competence,* Language, Speech and Communication Series, a Bradford Book. Cambridge, Massachusetts and London, England: The MIT Press.

Pavel, Silvia, and Diane Nolet. 2001. *Précis de terminologie*. Hull: Travaux publics et Services gouvernementaux Canada — Bureau de la traduction.

Picht, Heribert, and Jennifer Draskau. 1985. *Terminology: An Introduction*. Vol. 2: Surrey: University of Surrey, Department of Linguistic and International Studies.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company

Pozzi, María. 2001. "The Terminological Definition. Conflicts between Theory and Practice." *Language for Special Purposes: Perspectives for the New Millennium* 1:272-281.

Rondeau, Guy. 1984. *Introduction à la terminologie*. 2nd ed. Québec: Gaëtan Morin.

Rousseau, Louis-Jean. 1983. "La définition terminologique." In *Problèmes de la définition et de la synonymie en terminologie : actes du Colloque international de terminologie, Université Laval, Québec,* edited by Association internationale de terminologie, GIRSTERM, OLF, DGTD and Infoterm, 35-46. Québec: Girsterm.

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

Sager, Juan Carlos, and Marie-Claude L'Homme. 1994. "A Model for the Definition of Concepts : Rules for Analytical Definitions in Terminological Databases." *Terminology* 1 (2):351-373.

Sager, Juan Carlos, and Augustin Ndi-Kimbi. 1995. "The Conceptual Structure of Terminological Definitions and their Linguistic Realisations: A Report on Research in Progress." *Terminology* 2 (1):61-85.

Sager, Juan, and Kyo Kageura. 1994. "Concept Classes and Conceptual Structures : Their Role and Necessity in Terminology." *ALFA : Terminology and Special Linguistics* 7 (8):191-216.

Seppälä, Selja. 2006a. "Semi-Automatic Checking of Terminographic Definitions." In *Proceedings of the International Workshop on Terminology Design: Quality Criteria and Evaluation Methods (TermEval) — LREC 2006*, Genoa, Italy, 22-27.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

Seppälä, Selja. 2006b. “Vérification automatique de définitions terminographiques.” In *Journée d'étude de l'ATALA : Aide à la rédaction — Apports du Traitement Automatique des Langues*, Paris, http://perso.limsi.fr/amax/recherche/atala06/Seppala_atala06.pdf (July 25, 2014).

Seppälä, Selja. 2007. “Repérage automatique de génériques dans les définitions terminographiques.” In *Actes de la 11ème Recontre RECITAL*, Toulouse, 447-456.

Seppälä, Selja. 2009. “A Proposal for a Framework to Evaluate Feature Relevance for Terminographic Definitions.” In *Proceedings of the 1st Workshop on Definition Extraction*, Borovets, Bulgaria, 47-53.

Seppälä, Selja. 2010. “Automatiser la rédaction de définitions terminographiques : questions et traitements.” In *RECITAL 2010*, Montréal, Canada, http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_101.pdf (July 25, 2014).

Seppälä, Selja. 2012. *Contraintes sur la sélection des informations dans les définitions terminographiques: vers des modèles relationnels génériques pertinents*. Thèse de doctorat, Département de traitement informatique multilingue (TIM), Faculté de traduction et d'interprétation, Université de Genève.

Smith, Barry. 2003. “Ontology.” In *Blackwell Guide to the Philosophy of Computing and Information*, edited by L. Floridi, 155-166. Oxford: Blackwell.

Smith, Barry. 2004. “Beyond Concepts: Ontology as Reality Representation.” In *Formal ontology in information systems: proceedings of the third conference*

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

*(FOIS 2004)*, ed. By Achille C. Varzi and Laure Vieu, 73-84, Turin, November 4–6 2004. IOS Press.

Smith, Barry. 2008. “Searle and de Soto: The New Ontology of the Social World.” In *The Mystery of Capital and the Construction of Social Reality*, ed. by David Mark, Barry Smith and Isaac Ehrlich, 35-51. Chicago and La Salle: Open Court Pub Co.

Smith, Barry. 2011. “How to Do Things with Documents.” *Rivisti di Estetica* (preprint version):1-30.

Smith, Barry. 2012. “On Classifying Material Entities in Basic Formal Ontology.” In *Proceedings of the Third Interdisciplinary Ontology Meeting*, 1-13. Tokyo: Keio University Press.

Smith, Barry. 2014. “Document Acts.” In *Institutions, Emotions, and Group Agents. Contributions to Social Ontology*, ed. by Anita Konzelmann Ziv and Hans Bernhard Schmid, 19-31. Netherlands: Springer.

Smith, Barry, and Werner Ceusters. 2010. “Ontological Realism: A Methodology for Coordinated Evolution of Scientific Ontologies.” *Applied Ontology* 5:139-188.

Smith, Barry, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan Rector, and Cornelius Rosse. 2005. “Relations in Biomedical Ontologies.” *Genome Biology* 6 (5):R46.

Smith, Barry, and Pierre Grenon. 2004. “The Cornucopia of Formal-Ontological Relations.” *Dialectica* 58 (3):279-296.

Spear, Andrew D. 2006. *Ontology for the Twenty First Century: An Introduction with Recommendations*. Saarbrücken, Germany: Institute for Formal Ontology and Medical Information Science.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

Suonuuti, Heidi. 1997. *Guide to Terminology*, *NORDTERM 8*. Helsinki: Finnish Centre for Technical Terminology.

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

Temmerman, Rita. 2000b. "Une théorie réaliste de la terminologie : le sociocognitivisme." *Terminologies nouvelles* (21):58-64.

*The terminology database of the Swiss Federal Administration (TERMDAT)*: https://www.termdat.bk.admin.ch (May 29, 2014).

*TERMIUM Plus*®: http://www.btb.termiumplus.gc.ca (May 29, 2014).

*UBS Dictionary of Banking*: http://www.ubs.com/global/en/DictionaryOfBanking.html (May 29, 2014).

Vézina, Robert, Xavier Darras, Jean Bédard, and Micheline Lapointe-Giguère. 2009. *La rédaction de définitions terminologiques*, *Version abrégée et adaptée par Jean Bédard et Xavier Darras*. Montréal: Office québecois de la langue française.

Weissenhofer, Peter. 1995. *Conceptology in Terminology Theory, Semantics and Word-Formation : A Morpho-Conceptually Based Approach to Classification as Exemplified by the English Bbaseball Terminology*. Vienna: TermNet.

Zawada, Britta, and Piet Swanepoel. 1994. "On the Empirical Adequacy of Terminological Concept Theories: The case for Prototype Theory." *Terminology* 1 (2):253-275.

Author's address

Selja Seppälä

State University of New York at Buffalo

Department of Philosophy

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 

Park Hall, Buffalo, NY 14260-4150

seljamar@buffalo.edu

About the author

Dr. Selja Seppälä is a postdoctoral researcher at the University at Buffalo, where she has been working under the supervision of Dr. Barry Smith since 2012. Her research project, funded by the Swiss National Science Foundation (SNFS), aims at laying the groundwork for creating computer-assisted natural language definition writing tools leveraging ontological data that can be used in any possible context and practice, both by terminologists and ontologists. Her research interest is in definitions in dictionaries and in ontologies, and in the automation of definition processing, production and checking. Her primary domains of expertise are terminology and natural language processing with a strong interest in ontologies and philosophy.

Author accepted manuscript of:
Selja Seppälä, An ontological framework for modeling the contents of definitions. *Terminology* 21:1 (2015), 23-50. doi 10.1075/term.21.1.02sep.
issn 0929–9971 / e-issn 1569–9994 © John Benjamins Publishing Company