An Ontology-terminology Model for Designing Technical e-dictionaries: Formalisation and Presentation of Variational Data

Laura Giacomini

Hildesheim University/ Heidelberg University
E-mail: laura.giacomini@iued.uni-heidelberg.de

Abstract

This paper presents a model for the description of terms and term variants in technical e-dictionaries designed for professional translators and technical writers. The paper introduces a concept of variation as a phenomenon affecting (quasi)synonymous terms and terminological word combinations with morphological affinity, and provides an overview of the methodological steps involved in ontological/semantic systematisation, in morphosyntactic analysis of terminological variants and in the following data formalisation. The model is based on a multi-layered formalisation procedure that includes the compilation of a coherent domain ontology, the identification of domain-specific frames and frame elements, and the description of term variants along a previously designed morphology-oriented typology. The paper also discusses visualisation options and search query types in the final e-dictionary. Examples are taken from German and English terminology related to thermal insulation products, with the purpose of hinting at the general applicability of the model to other technical subfields.

Keywords: terminology; ontology; variation; technical domain; LSP; e-dictionary

1. Introduction

This paper presents a description model for terms and term variants in technical e-dictionaries. The study is part of a larger project on database representation of terminological variation, in which restricted technical subdomains, belonging to the areas of building and electrical engineering, have been analysed and compared to test the feasibility of the method. Despite clear differences at the level of conceptualisation, standardisation and communicative features between the two domains, the model has proven to be globally efficient, and seems to provide a reliable method for conceptual and terminological representation in other comparable technical subfields. The employed method and the resulting lexicographic presentation are explained via reference to German and English terms belonging to the field of building thermal insulation. First, ontological data are introduced together with their descriptors (Section 2.1). Second, lexical data (terms and variants) are classified along morphosyntactic rules (Section 2.2). In the next step, the method for merging ontological and morphosyntactic formalisation is discussed. It is also shown how conceptual and lexical formalisation can be embedded in NLP procedures for extracting candidate terms from LSP corpora (Section 2.3). The concluding part of the paper concentrates on visualisation features of the final lexicographic product (Section 3).

2. Systematic analysis and multi-layered formalisation

2.1 Ontological and semantic data analysis

At the core of conceptual formalisation is a domain ontology and its association with a frame-based approach to obtain fine-grained data descriptions. The domain ontology has been built on the grounds of knowledge that was manually retrieved from several specialised texts dealing with the topic of thermal insulation. These texts, which may address different target recipients, belong to the most typical genres in this field, e.g. specialised magazines, handbooks, product descriptions, data sheets, and guides. Due to the complex structure of the ontology and, in particular, its integration with frame elements, a formalisation of ontological knowledge and the corresponding lexical information by means of widespread RDF models (e.g. lemon-OntoLex) has been avoided, at least for the moment. This type of ontological and semantic data has been recorded in a relational database in the same way as terminological data, rather than in a database-external conceptual layer (as is the case of Ontop and similar OntoLex systems, cf. Bosque-Gil et al., 2015).

The domain ontology is structured around a key entity (or class of entities), the THERMAL INSULATION PRODUCT(S), which constitutes the topical focus of a collection of reference texts[1]. It consists of objects and their taxonomic and non-taxonomic relationships (Declerck & Gromann, 2012). Taxonomies may regard both instances and classes of instances, and produce a controlled vocabulary with a hierarchical structure of the kind parent-child or superclass-subclass. Ontological knowledge representation, however, often requires other types of information to express relations between entities as well as properties of entities. As for first-order entities (Lyons, 1977) in the form of inanimate objects, it is useful to approach ontology work by employing a tripartite supercategorisation as a starting point: thermal insulation products can be observed by taking into consideration aspects regarding their MATERIAL, their FORM and their FUNCTION. Each of these macrocategories includes a number of entities that are sometimes taxonomically related to other entities of the ontology. For instance, a category that is connected to the function of thermal insulation products is the BUILDING COMPONENT to which the product is applied, whereby a specific building component is a kind of superordinate entity belonging to a building component class (e.g. a flat roof is a kind of roof, cf. Table 1).

---

[1] Texts have been selected on the grounds of their relevance for translation (typically translated texts concerning this topic) and for companies (typically published texts concerning manufacturing, selling and application of a specific product).

BUILDING COMPONENT	CLASS OF BUILDING COMPONENTS
flat roof	roof
mono-pitched roof	roof
multi-pitched roof	roof
exterior wall	wall
interior wall	wall

Table 1**:** Example of an ontological category with taxonomic relationships.

Other entities belong to ontological categories in the form of terms that do not build taxonomic relationships to other relevant categories. For instance, PRODUCT USER and PRODUCT FEATURE, both belonging to the macrocategory FUNCTION, show this kind of behaviour (cf. Table 2).

PRODUCT USER
technician/craftsman
handyman

PRODUCT FEATURE
fire-resistance rating
thermal conductivity
heat storage capacity
bulk density

Table 2**:** Examples of ontological categories without taxonomic relationships.

Ontological categorisation is integrated with more specific semantic information in the form of frame elements in terms of the Frame Semantics theory. The key entity, the THERMAL INSULATION PRODUCT, is seen as part of one of the potential frames involving that entity. Frames are typical situational perspectives, in which specific entities (frame elements) play a role. For instance, thermal insulation products (the concept and the corresponding terms) can be considered from the perspective of their production, their selling, or their use. In the preferred frame, in this case, the insulation product is an object with distinctive features that is sold by producers or traders to specific users in order for them to thermally insulate one or more components/areas of a building. The selected frame serves as an interface between the ontology and the lexicon of the subdomain, and provides a relevant tool for semantic categorisation of terms as well as for lexicographical disambiguation of variants.

Each term or term component directly denoting or indirectly referring to a thermal insulation product can be reduced to a frame containing all or some of its typical frame elements. These elements, e.g. MATERIAL, FORMAT, PART OF THE BUILDING, APPLICATION TECHNIQUE, can be understood as potential semes which coincide with the previously identified ontological entities. The relationship between the ontological, semantic and terminological levels of the proposed model can be visualised as follows:

ontological level:	ontological (macro)category	FORM > FORMAT	FUNCTION > APPL. TECHNIQUE
semantic level:	frame element	FORMAT	APPL. TECHNIQUE
term level:	terms/variants	*insulation boards, insulation batts*	*spray foam insulation, blow-in insulation*

The domain ontology should be a general (if not universal) picture of the objects/concepts that compose the domain, whereas the chosen frame is embedded in the description of a specific situation and, accordingly, can match different constellations of ontological entities. At the terminological level, single terms and multiword terms can be subdivided into semantic components that are directly related to the elements of the relevant frame.

2.2 Terminological data analysis

The study concentrates on terminological variation as a key phenomenon in specialised language, which, in recent decades, has been analysed along different theoretical approaches (cf., among others, Auger, 2001; Freixa, 2006). Our description of terminological variants is based on a concept of variation as a phenomenon affecting (quasi)synonymous terms and terminological word combinations with morphological affinity (i.e. they share at least one lexical morpheme). Texts concerning thermal insulation products, belonging to different textual genres and embedded in various communicative situations, often include more or less large clusters of semantically and morphologically homogeneous terms; for instance, *wood fibre insulation boards, wood fiber insulation boards, wood fibre thermal insulation boards, wood fibre boards for acoustic and thermal insulation, wood fibre boards for external wall insulation, wood fibre boards for insulating walls internally and externally*, etc. The generally low degree of standardisation in the subfield of thermal insulation, in which international and national standards provide guidance and specifications only for a part of the involved entities, is one of the main causes of intensive variation. Variant clusters are apparently relevant in technical writing and specialised translation, but language professionals in these fields are often compelled to perform time-consuming queries in parallel and comparable corpora to obtain information on the availability and correctness of potential variants. Lexicographic information tools such as LSP e-dictionaries, glossaries and termbases, in fact, cover only a small fraction of the commonly used variants. They usually record possible variants at different levels of discourse, for instance geographical variants such as *fibre* (BrE) and *fiber* (AmE) or, in general, variants with no morphological affinity (e.g. German *isolieren/ dämmen*), which, however, are relatively infrequent in specialised language. On the contrary, morphologically similar synonyms at the same level of discourse are not taken into consideration, with the exception of rare cases. Variants extracted from texts are assigned to classes according to a language-independent variation typology. Each synonymous variant of a term is classified along morphological, syntactic and graphical criteria. Graphical variation regards phenomena such as hyphenation, and plays a minor role in variation analysis, not least because these phenomena are scarcely subject to standardisation[2]. Morphological variation may be total, partial, or entirely missing. Figure 1 shows the three most relevant combinations of variation types, which correspond to the light grey areas.

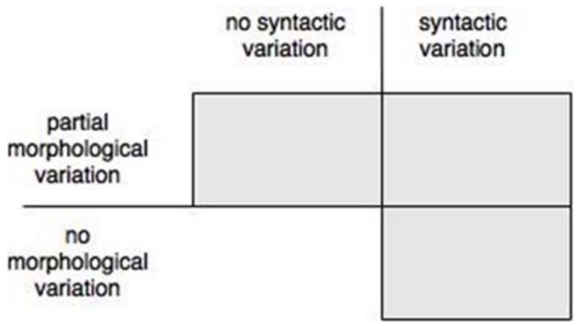


Figure 1: Relevant types of term variation

Since the study concentrates on morphologically similar variants, the focus of the study lies on partial morphological variation, independent of its combination with syntactic changes, as well as on syntactic variation without morphological change (light grey areas). Morphological change is missing whenever the variant of a term is made of the same lexical morphemes as the original term.

Variation types indicated in Figure 1 can be illustrated by means of the following examples in German and English[3]:

a) partial morphological variation and no syntactic variation (pMV-nSV)

    DE: *Dämmstoff/Isolierstoff*

    EN: *thermal insulation/heat insulation*

b) partial morphological variation and syntactic variation (pMV-SV)

    DE: *Dämmstoff/wärmeisolierender Stoff*

    EN: *polystyrene foam insulation/insulation with styrofoam*

c) no morphological variation and syntactic variation (nMV-SV)

---

[2] Referring to Figure 1, it can be noted that a case of non-morphological and non-syntactic variation could coincide with mere graphical variation, for instance the absence or presence of hyphenation in the two German terms *Polyurethanschaum/Polyurethan-Schaum* (EN *polyurethane foam*).

[3] For simplification reasons, the following abbreviations have been assigned to the relevant variation classes and types: MV = morphological variation; pMV = partial morphological variation; nMV = no morphological variation; SV = syntactic variation; nSV = no syntactic variation.

DE: *Isolierung der Fenster/Fensterisolierung*

EN: *roof insulation/insulation of roofs*

Terms belonging to clusters of morphologically similar synonyms can be analysed on the basis of the presented variation types. Variation can be classified either confronting terms pairwise or, as far as a preferred term can be identified along an existing standard or by conventional use, referring available variants to the preferred term. The following two examples illustrate both approaches to classification:

	pairwise relations:	relations to the preferred term:
(A)		
wood fibre insulation boards (preferred term)		
	} pMV-SV	
wood fiber insulation boards		pMV-nSV
	} pMV-SV	
wood fibre thermal insulation boards		pMV-SV
	} nMV-SV	
wood fibre boards for thermal insulation		pMV-SV
	} pMV-SV+	
wood fibre boards for external wall insulation		pMV-SV+
	} nMV-SV	
wood fibre boards for insulating walls externally		pMV-SV+
(B)		
WDVS-Mineralwolle (preferred term)		
	} nMV-SV	
Mineralwolle als WDVS		nMV-SV
	} pMV-SV	
Mineralwolle zum Dämmen im WDVS		pMV-SV
	} pMV-nSV	
Mineralwolle zum Dämmen im Wärmedämmverbundsystem		pMV-SV
	} nMV-SV	
Mineralwolle... einer der beliebtesten Dämmstoffe für WDVS		pMV-SV

The "+" sign in the first example indicates a semantic expansion: underlined words (cf., for instance, *external wall*) add conceptual information to the contrasted original term, automatically changing both its morphological and semantic nature.

Pairwise classification enables fine-grained interpretation without postulating a hierarchical structure between a preferred term and its variants. In textual analysis, this approach can be useful to follow up variation strategies and motivations inside a specialised text. From a lexicographic perspective, however, classifying variants by comparing them to a preferred term has more advantages, since it may enable a dictionary user to retrieve all variants of a lemma belonging to a specific type with the help of variation-related filters (cf. Section 3).

2.3 A multi-layered formalisation: ontological, semantic and terminological data for the lexicographic database

Data formalisation takes place with the help of ontological, semantic and variational descriptors that are meant to provide lexicographic users with comprehensive information concerning terms and variants of the selected technical domain. Given a term and its synonymous variants, the formalisation process can be illustrated using the examples of *stone wool insulation batts* and *Holzfaserdämmplatten* (*wood fibre insulation boards*), and their synonymous variants (Table 3).

Common terms (i.e. non-proper nouns) indicating a thermal insulation product can be decomposed in semantic unities that refer to specific frame elements and occur in all synonymous variants in order to produce the same term meaning. In the exemplified case, the three frame elements MATERIAL (insulation material of which the product is made), FORMAT (the format in which the product is sold) and GOAL (the purpose with which the product is employed) constitute the semantic profile of the preferred terms *stone wool insulation batts* and *Holzfaserdämmplatten*, and their variants. Variants can combine these frame elements in a different syntactic order:

stone wool insulation batts vs. *insulation batts made of stone wool*

and/or introducing morphological transformations:

stone wool vs. *mineral wool*

insulation vs. *thermal insulation*,

building in this way a large cluster of multiword terms which share the same semantic head, *batts*. When compared with the preferred terms, variants display heterogeneous combinations of morphological and semantic changes.

The lexicographic database, which is structured in a relational form, records in its tables, for each type of thermal insulation product,

- the preferred term indicating this product and

- its terminological variants (synonyms);
- for each variant, the involved variation type;
- the semantic decomposition of the preferred term and its variants[4];
- the frame elements which are realised by the preferred term and its variants.

Frame elements	Terms	Variation	
	[stone wool] [insulation] [batts] (pref. term)	MV	SV
MATERIAL = [stone wool] FORMAT = [batt] GOAL = [insulation]	[mineral wool] [insulation] [batts]	partial	-
	[stone wool] [thermal insulation] [batts]	partial	✓
	[stone wool] [batts] for [thermal insulation]	partial	✓
	[stone wool] [batts] for [insulating] ...	-	✓
	[insulation] [batts] made of [stone wool]	-	✓
Frame elements	Terms	Variation	
	[Holzfaser][dämm][platten] (pref. term)	MV	SV
MATERIAL = [Holzfaser] FORMAT = [Platte] GOAL = [Dämmung]	[Holzfaser][platten] zur [Dämmung] ...	partial	✓
	[Dämm][platten] aus [Holzfasern]	-	✓
	aus [Holzfasern] hergestellte [Dämm][platten]	-	✓
	[Platten] aus [Holzfasern] zur [Dämmung] ...	-	✓

Table 3 – Example of data formalisation by means of frame elements and variation types.

Semantic decomposition, signalled in Table 3 by means of square brackets, is essential for the identification of ontological/semantic differences and similarities between terms that possibly embody other frame element constellations. The frame element FORMAT, for instance, is realised in English by terms such as *slab, board, mattress, rope, foam*, or

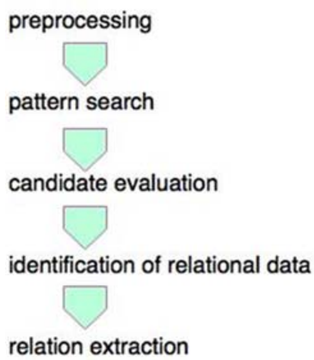
[4] A morphosyntactic decomposition of terms is also provided. However, this topic will not be discussed in this paper.

loose granules, which may belong to larger multiword units together with terms indicating other frame elements. This means that the same frame elements combination can be found in several terms, depending on the language-independent finite number of relevant ontological entities, as well as on the language-specific availability of synonyms (e.g. *panel/ board*), as shown in this example:

	[mineral wool] [mattress]
	[fibreglass wool] [mattress]
	[mineral wool] [batt]
MATERIAL + FORMAT:	[mineral wool] [slab]
	[polystyrene] [panel]
	[polystyrene] [board]
	[polystyrene] [granules]
	[perlite] [granules]

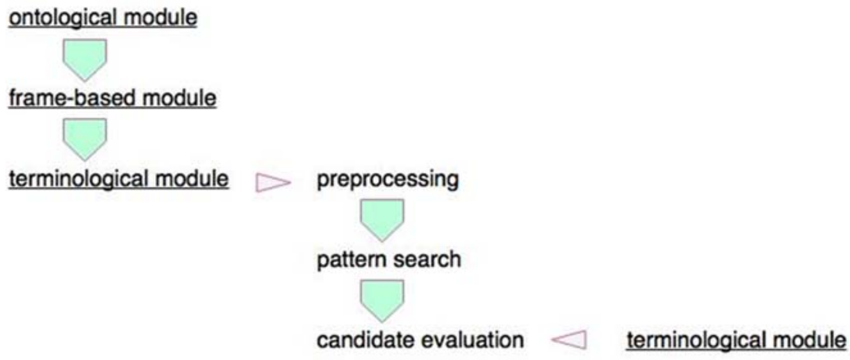
The proposed model of lexical data representation could be combined with NLP techniques to term and relation extraction from LSP corpora to create a semi-automatic pipeline for improving identification of semantically related terms. The formalisation process, as a matter of fact, provides the basis for a consistent rule-based morphosyntactic and semantic analysis, with a direct connection between the two analysis levels. Existing NLP procedures aimed at relation extraction (cf. Rösiger et al., 2016) are based on an inductive method, i.e. on specific instances that lead to generalised statements: relational data obtained by means of corpus preprocessing, pattern search and candidate evaluation are used to extract further relations (a).

(a)



A procedure which integrates the proposed formal model consisting of the ontological, the frame-based and the terminological (morphosyntactic, variational) module allows for the preliminary annotation of terms with new ontological, semantic and morphosyntactic information and thus for a deductive identification of relational data. The new descriptive modules (cf. underlined elements) can interface in different ways with the basic pipeline, depending on the process stage in which their information (sense, morphosyntactic, and variation tags) is most required (b).

(b)



In example (b), terminological formalisation could take place both at a preprocessing and evaluation level, for enhancing both precision and recall of retrieved data. The process could also incorporate bootstrapping techniques in order to iteratively refine extraction results for both terms and relations in a corpus. Extracted candidate terms need to undergo evaluation, preferably in a semi-automatic procedure in which linguistic and conceptual coverage are tested, with the previously identified terminological profiles potentially serving as one of the non-automatic validation tools (for a comparable approach cf. Christensen, 2016).

3. Data presentation in the technical e-dictionary

Terms concerning thermal insulation products are recorded in the technical e-dictionary together with other lexical data (e.g. variants, equivalents in a target language, usage examples) and metadata (e.g. frame elements and relevant entities from the underlying ontology). The main visualisation features in the dictionary will now be shown and discussed (cf. Costa, 2013). These include different presentation modes for conveying user-tailored lexicographic information. Ideally, target users should in fact be able to select specific information, i.e. to group variants along morphosyntactic or conceptual/semantic principles by applying more or less detailed filters.

The addressed user is the technical writer and the professional translator passively translating into the native language. Lexicographically relevant needs arising in specific extralexicographic situations (Tarp, 2008) determine specific dictionary functions to efficiently provide potential users with the required assistance. Technical writers produce functional and user-oriented specialised texts, particularly technical documentation (Göpferich 1998: 1003), whereas professional translators need to produce a native-language target text being tied to a foreign-language source text. Despite this operational and cognitive difference between the two tasks, the presented model aims to serve both target groups by means of a clear text-productive orientation: the main function of the technical e-dictionary is to make variants, and information about variants, available to users in their native language, independent of the qualitative and quantitative features of variation in foreign-language reference resources used for technical documentation or in a foreign-language source text that has to be translated. On the one hand, dictionary users should be able to perform separate or combined queries involving each data type available in the database. On the other hand, they should be able to customize a search by applying filters to single data types in order to obtain tailored results.

Table 4 displays some of the manifold possibilities of performing search queries by combining different levels of knowledge about variational data. For instance, example (b) given for query type (2) is a combination of query type (2), i.e. the search for a specific variational profile, with query type (1), i.e. the search for a term or part of a term.

Query types, specifically (1)–(3), refer to a terminological layer including terms, variants and their variational and morphosyntactic description. Query types (4)–(5) are related to the ontological/semantic layer of the database, with information concerning frames and the domain ontology. The structure of the database and the multi-layered data formalisation allow for targeted search queries and the combination of query types during a single search act. Users can choose whether to look up a term or to start a query by indicating, for instance, a well-defined set of frame elements, or even if they wish to combine both kinds of information to obtain more fine-grained results (cf. possible query relations in the second column of Table 4).

At the same time, filtering as well as result-widening options in the form of expand/hide commands can be selected during each search query in order to retrieve either more specific or more general results. For instance, the output of the first query example would include by default the term, its classified variants, their morphosyntactic structure, usage examples, as well as the involved frame elements and ontological categories (a). This also constitutes the microstructure of lexicographic entries. However, users can also choose to expand on further results that include additional frame elements (d).

Against the background of the specific user's needs and the relevant microstructural items, it is clear that the technical e-dictionary has both a form-determined and a systematic macrostructure, and that it allows for multiple access paths to the desired data (cf. Giacomini, 2015). Moreover, data representation in the lexicographic database allows for both a monolingual and a bilingual coverage of terms and variants.

Search query type:	Query relations:	Query example (with specific query relation):	Output example:
(1) search for a single or (part of) multiword term		flat roof insulation	(a) preferred term: flat roof insulation variants: flat roof thermal insulation, insulation for flat roofs, insulating flat roofs + *variants types* + *morphosyntactic types* + *information on involved frame elements and ontological entities*
(2) search for a variation profile	and (1)	and (1) nMV-SV	(b) insulation for flat roofs, insulating flat roofs
(3) search for a morphosyntactic structure	and/or (1), and (4)(5)	and (1) NN	(c) preferred term: flat roof insulation variants: flat roof thermal insulation, insulation for flat roofs
(4) search for (a) (combination of) frame element(s)	and/or (1), and (3)(5)	and (1) + MATERIAL + FORMAT	(d) WOOD FIBRE BOARDS for flat roof insulation, EXPANDED POLYSTYRENE SLABS for flat roof insulation
(5) search for an ontological entity or category	and/or (1), and (4)	and (1) MATERIAL	(e) [WOOD FIBRE]/ [POLYSTYRENE]/ [POLYURETHANE] + flat roof insulation

Table 4 – Search query types and visualisation options in the technical e-dictionary.

4. Conclusions

This paper has introduced a description model for technical terms and their variants in an e-dictionary designed for professional translators and technical writers, and covering terminology related to thermal insulation products. The aim of the paper was, on the one hand, to provide an overview of the methodological steps involved in ontological/semantic systematisation, in morphosyntactic analysis of terminological variants and in the following data formalisation. On the other hand, a major goal of this paper was to discuss visualisation options in the final e-dictionary, and to associate them with its overall microstructural, macrostructural and access properties.

As already mentioned in the introduction, results presented in this paper, as well as those obtained in the underlying project concerning this topic, confirm the effectiveness of the method in:

- creating multi-layered, language-independent descriptions for synonymous variation in restricted technical subdomains;

- adapting this description to lexicographic functions of resources that address specific target users; and

- providing formalisation tools to possibly improve NLP procedures for term and variant extraction from specialised corpora.

In the current project, synonymous variation and its morphological peculiarities are at the centre of discussion as one of the most pervasive and, at the same time, underestimated lexical phenomena in terminology. Its relevance for LSP dictionaries addressing professional text producers is indisputable. Special attention is due in the field of electronic lexicography, which can provide the necessary tools (e.g. data formalisation, or database representation strategies) to ensure extensive, modular coverage of the phenomenon, and which can benefit from the availability of data obtained by semi-automatic term and variant extraction. Future work conducted on the language(s) of technology will further investigate these topics and attempt to expand the method of technical (and maybe non-technical) subdomains displaying even larger differences in conceptualisation, standardisation and communicative features.

5. References

Auger, P. (2001). Essai d'élaboration d'un modèle terminologique/terminographique variationniste. In *TradTerm*, *7*, 183-224.

Bosque-Gil, J. et al. (2015). Applying the OntoLex model to a multilingual terminological resource. In *European Semantic Web Conference.* Springer International Publishing, 283-294.

Christensen, L. W. (2016). Semi-automatic Evaluation of Terminological Web-crawled Corpora. *Term Bases and Linguistic Linked Open Data*, 64.

Costa, R. (2013). Terminology and Specialised Lexicography: two complementary domains. In *Lexicographica* 29.1/2013, 29-42.

Declerck, T. & Gromann, D. (2012). Combining three ways of conveying knowledge: Modularization of domain, terminological, and linguistic knowledge in ontologies. In *Proceedings of the 6th International Workshop on Modular Ontologies, Graz, Austria, CEUR-WS, Aachen. CEUR Workshop Proceedings* Vol. 875, 28-40.

Fillmore, C. J. (1977). Scenes-and-frames semantics. In A. Zampolli (ed.) *Linguistic Structures Processing.* Amsterdam: North-Holland Publishing Company, 55-81.

Freixa, J. (2006). Causes of denominative variation in terminology: A typology proposal. In *Terminology* 12(1). Amsterdam: John Benjamins, 51-77.

Giacomini, L. (2015). Macrostructural properties and access structures of LSP e-dictionaries for translation. In *Lexicographica* 2015/31. Berlin/Boston: De Gruyter. 90-117.

Göpferich, S. (1998). Schreiben in der Technik/Technical Writing. Fachsprachen: Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. New York, Berlin: de Gruyter, 1003-1014.

Lyons, J. (1977). *Semantics* (vols. i & ii). Cambridge CUP.

Rösiger, I. et al. (2016). Acquisition of semantic relations between terms: how far can we get with standard NLP tools? In *Computerm 2016*, 41.

Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-Knowledge. Tübingen: Max Niemeyer Verlag.