Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary

Laura Giacomini
Heidelberg University
E-mail: laura.giacomini@iued.uni-heidelberg.de

Abstract

In this paper, a frame-based approach to terminological variation is presented along a model for presentation of multiword terms and their variants in a technical e-dictionary. A case study concerning terminology related to semiconductor diodes is the background against which methods and goals of a larger study on the technical language (Habilitation thesis at Hildesheim University) are illustrated and compared with those of existing resources. At the core of the proposed model are three interrelated description layers (ontology – frame – terminology), with the frame layer serving as the semantic interface between ontological classes and terms, as well as a variation typology accounting for orthographical, morphological and syntactic term variants. The microstructural properties of the envisaged e-dictionary, which aims at supporting text production in the native language, are illustrated by means of the example entry *diode in forward bias*. The addressed users, technical writers and professional translators, are able to access all types of data separately from each other, in a modular way. The paper closes with an outlook on how future developments could include application of the model to further technical domains.

Keywords: frame-based terminology, frame-based lexicography, term variation, technical language, LSP dictionary

1 Introduction

This paper describes a model for data presentation in a technical e-dictionary with special focus on variation of multiword terms. This is part of a larger corpus-based study on the modelling of a terminological database for lexicographic purposes[1].

A multiword term is functionally understood as "a term containing two or more content words" (Jacquemin & Tzoukermann 1999: 26). The technical e-dictionary, which aims to fill a clear gap in lexicographic coverage of variation, is intended to support text production and specifically addresses professional translators and technical writers. The proposed data representation method and the related lexicographic presentation are explained by using English denominations employed in the field of electrical engineering and referring to semiconductor diodes (Figure 1). Generally speaking, semiconductors are "solids whose electrical conductance lies between that of good conductors and insulators" (Clouden 2014: 80).

A diode is a specialized electronic component with two electrodes called the anode and the cathode. Most diodes are made with semiconductor materials such as silicon, germanium, or selenium. […] Diodes can be used as rectifiers, signal limiters, voltage regulators, switches, signal modulators, signal mixers, signal demodulators, and oscillators. The fundamental property of a diode is its tendency to conduct electric current in only one direction. (http://whatis.techtarget.com)

1 Habilitation thesis at the Institute of Information Science and Natural Language Processing, Hildesheim University (Germany).

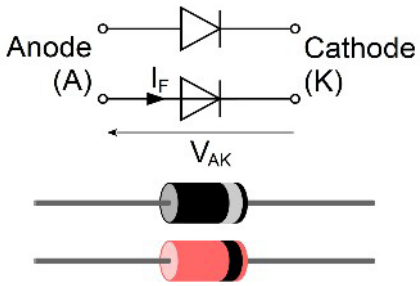


Figure 1: Semiconductor diode and diode symbol with indication of the anode and cathode.[2]

The terminology of electrical engineering is subject to a comparatively high degree of standardization, with national and international institutions and associations contributing to unification of terminology in its different subject areas (e.g. electromagnetism, circuit theory, and computer network technology, to name just a few that are part of the Electropedia classification; cf. http://www.electropedia.org). In comparison with other technological areas, this results in a lexically homogeneous domain. Still, synonymous variation appears to be a quantitatively and qualitatively relevant phenomenon which deserves careful attention. So far, many studies in computational linguistics (cf. Daille 2005, Daille 2017), terminology and translation (cf. Fernández-Silva & Kerremans 2011, Temmerman 2000) have systematically explored terminological variation, conveying a view of terminology that is quite different from its more traditional, monolithic interpretations.

After pointing out which notion of variation and which types of variants will be taken into consideration (Section 2), we will concentrate on a frame-based approach to data modelling (Section 3), and then finally on data presentation in a technical e-dictionary (Section 4). The last section of this paper recapitulates key findings and draws some conclusions on the significance of frame-based specialized lexicography and the applicability of the proposed description model to other technical domains.

2 Term Variation

2.1 Which Notion of Variation?

In the context of this study, terminological data is retrieved from a database that collects terms, variants and relations extracted from corpora of specialized texts and addressing experts and semi-experts. Multiword terms are at the center of discussion as terms which are most exposed to non-diasystemic variation, i.e. to (near) synonymous variation occurring at the same discourse level with no diachronic, diatopic, register-, or corporate language-related changes (Caro Cedillo 2004), e.g.

depletion region
depletion zone
depletion layer
space charge region
space charge layer.[3]

The study focuses on variants which match the following criteria:

2 Erik Streb CC-BY-SA-3.0 http://en.wikipedia.org/wiki/File:Diode_pinout_en_fr.svg

3 "Near the junction, a depletion region is created by electrons from the N-type material moving in to fill holes in the P-type material, and holes moving in the opposite direction (from the P-type material) to combine with available electrons. The depletion region is electrically neutral, but separates the N- and P-type materials, which have a difference in potential called the barrier potential (or junction voltage)" (Diffenderfer 2005: 41).

- Variants are totally or partially synonymous,
- Variants display a morphological similarity (similarity is hereby defined as the presence of shared lexical morphemes), and
- Variants build term clusters which mostly belong to the same systemic level. Diasystemic, e.g. geographical, variation is still accounted for if available, but does not represent the focus of this study.

Empirical observation of term behavior in technical language often provides evidence of variant clustering in the same text, with the use of variants motivated by discourse-related, functional, interlinguistic, and cognitive factors (for a detailed coverage of variation grounds, cf. Freixa 2005). Here are a few examples of synonymous variation within the same text:

> "a diode… is forward biased […] the biasing is classified as Forward biasing and Reverse biasing of a diode […] a diode is connected in a forward bias" (Godse & Bakshi 2010: 73)
>
> "The shape of the charge density, p, depends upon how the diode is doped. Thus, the junction region is depleted of mobile charge carriers. Hence, it is called the depletion region (layer), the space charge region, or the transition region. The depletion region is of order 0.5 pm thick. There are no mobile carriers in this very narrow depletion layer. Hence no current flows across the junction and the system is in equilibrium" (Salivahanan et al. 1998: 88)
>
> "SEMICONDUCTOR DIODE AS A RECTIFIER Figure 7.15 depicts the rectifying action of a semiconducting diode. […] In this way, the semiconductor diode has been able to do rectification, i.e., change ac into dc" (Joshi 2010: 7.12).

Moreover, the availability of a relatively large number of n-gram variants in various technical fields suggests that this is likely to be a common phenomenon in technical language. The presence of different degrees of extension due to field-specific properties does not refute these findings, it simply proves that variation is a natural process, at least in some LSP and that it is strictly interconnected with language- and text-dependent factors, for instance the type of communication involved (i.e. domain-internal or domain-external), and the specific register or source features.

2.2 Which Types of Variants?

The overall study deals with variant description at the single term and multiword term levels. Multiword terms, on which this paper concentrates, include both complex terms and phrasemes. We distinguish the following three types of variation:

1) MV, morphological variation (partial / total):
   changes in lexical morphemes (e.g. *depletion layer* vs. *depletion region*)
2) SV, syntactic variation:
   changes in the part of speech, word order, and sentence construction (e.g. *depletion region* vs. *region of depletion*)
3) OV, (ortho)graphical variation:
   changes in hyphenation and capitalization (e.g. *light-emitting diode* vs. *light emitting diode*).

In addition to the examples just mentioned, the three types often combine with each other, building complex patterns of variation.

3 Frame-based Data Modelling

In the present study, variant modelling is largely based on a frame-based approach to terminology, in which basic ideas deriving from Frame Semantics (Fillmore 1977, Ruppenhofer et al. 2006) and

Frame-based Terminology (Faber 2015) are adapted to the modelling of specialized discourse, with the purpose of representing terms of a certain domain according to the role they play in certain domain-specific scenarios. Some frame-oriented lexicographic and terminographic resources have already been published over the last ten years. Well-known examples are the multilingual EcoLexicon (Reimerink & Faber 2009), developed at the University of Granada and covering environmental terminology and the Kicktionary (Schmidt 2014), which deals with the language of football. The focus of our model with respect to the existing resources primarily lies in

- the inclusion of an extensive domain ontology which interfaces with the lexicon through the frame layer; the three layers of analysis (ontology – frame – terminology) are linked to and motivated by each other;
- the focus on terminological variation, with frame elements serving as identifiers of shared or distinct semantic roles in orthographical, morphological, and syntactic variants;
- the monolingual orientation of term and variant representation for supporting text production in the native language.

Data representation in the terminology database relies on a multi-layered model in which terms and variants undergo a top-down analysis process beginning with their conceptual background (ontological layer), going through their semantic content (frame layer) and ending with their morphological and syntactic features (see Table 1).

Table 1 – Multi-layered model and related procedural steps

LAYER	PROCEDURE	COMPONENTS
Domain ontology layer	Designing an ontology for semiconductor devices.	The ontology includes taxonomic and non-taxonomic relations between classes.
	Selecting a key ontological entity: SEMICONDUCTOR DIODE.	
Frame layer	Identifying possible frames related to the semiconductor diode, e.g. PRODUCTION or SALE.	
	Selecting a frame to be described in the model: FUNCTIONALITY.	The frame includes core and non-core frame elements.
Lexical layer	Modelling single terms, multiword terms and term variants.	Morphosyntactic, frame-related, ontological and variational features.

The top level of the proposed model is a domain ontology structured around a key entity, the SEMICONDUCTOR DIODE, which constitutes the topical focus of the available corpus texts. At the interface between the top ontological level and the bottom lexical level is a frame level in which the key entity is semantically accounted for in the sense of Frame Semantics (Fillmore 1977, Ruppenhofer et al. 2006) and Frame-Based Terminology (Faber 2015). The frame FUNCTIONALITY is selected among the possible frames describing a semiconductor diode, and each term or term component directly denoting or indirectly referring to a diode can be reduced to the identified frame elements (e.g. SEMICONDUCTOR MATERIAL, CONSTRUCTION FORM, APPLICATION TECHNIQUE). Figure 2 shows the combination of the three frame elements PRODUCT (PROD), GOAL (GOAL) and PROPERTY (PROP) in a set of synonymous multiword terms.

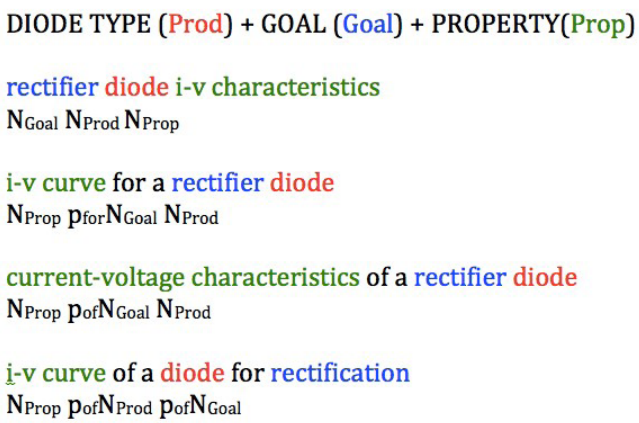


Figure 2 : Set of synonymous variants and corresponding syntactic-semantic annotation.

The frame-based approach to terminology, with its clear connection to cognitive linguistics (cf. Faber 2012), is at the core of the illustrated model. On the one hand, this approach establishes a link between the lexical and the ontological level, with frames seen as seme subsets and semes as semantic roles attached to ontological entities. On the other hand, this approach provides the necessary key for interpreting and describing the correspondence between morphosyntactic and semantic features of any multiword term.

The bottom level of the model envisages lexical analysis along morphosyntactic, conceptual (i.e. ontological and frame-related) and variational parameters, each supporting a different task in text production.

Any multiword term,

e.g. *rectifier diode i-v characteristics*,

in which the abbreviated form *i-v* stands for *current-voltage*, can be formally described in terms of

I its morphosyntactic structure: AP (A *rectifier* + N *diode*) + NP (N *i-v* + N *characteristics*),
II its rule-based relation to the basic morphosyntactic structures of its language: AP + NP,
III the frame elements denoted by its constituents: GOAL + PRODUCT + PROPERTY, and
IV the ontological classes to which these constituents are linked: FUNCTION: APPLICATION + SEMICONDUCTOR DIODE + MATERIAL: PROPERTY (see Table 2 for an overview of description I. to IV.).

Table 2: Term description based on the multi-layered model.

Domain ontology layer	FUNCTION: APPLICATION	SEMICONDUCTOR DIODE	MATERIAL: PROPERTY
Frame layer	GOAL	PRODUCT	PROPERTY
Lexical layer	A *rectifier*	N *diode*	N + N *i-v characteristics*

Furthermore, each variant of the given multiword term,

e.g. *i-v curve of a diode for rectification*,

is assigned to

VI. a specific variant class and type: partial morphological variation (*characteristics* > *curve*) + syntactic variation (AP NP > NP PP), and
VII. a specific variation template: paraphrase (*diode characteristics* > *curve of a diode*, *rectifier diode* > *diode for rectification*) + explicitation (*characteristics* > *curve*) + transposition (*rectifier* > *rectification*).

Variation is always identified in relation to a main term. The selection of a main term takes place by referring to existent standards and/ or to quantitative analysis in the available texts. However, the designation of a multiword term as a main term to which one or more variants are attached is a topic dealt with in the main study.

Frame-based data modelling has been developed, which pays special attention to the granularity of information. Descriptors, for instance frame elements, need to be specific enough to deliver a precise, unambiguous semantic characterization of terms, and general enough to be applicable to other technical domains. In particular, the feasibility of the model has been tested in other domains centered on technical artefacts (thermal insulation products and DIY-tools).

4 Data Presentation in a Technical E-dictionary

The parameters described in this section are the main building blocks of the abstract microstructure of a dictionary entry and will be discussed with the help of representative examples. The purpose of these examples is to comprehensively illustrate the model for lexicographic data presentation together with corresponding search options (semasiological and onomasiological access structures) and visualisation options.

Terminological variation is a phenomenon text producers have to deal with. The issue about the availability or adequacy of a given variant for a given context is well known among translators and technical writers. However, doubts cannot be removed by just using lexicographic resources, as lexicographic information tools (LSP dictionaries, glossaries and terminological databases) usually have the following characteristics:

- They cover only a small fraction of the commonly used variants;
- They rarely record longer multiword terms than bigrams;
- They usually contain possible variants at different levels of discourse (for instance geographical or chronological variants, e.g. *PNPN diode* vs. *Shockley diode*) or, in general, variants with no morphological affinity (e.g. *bias/ direction*);
- Their presentation produces coherency issues at macrostructural, microstructural and mediostructural level (variants may be lemmatized or not, may have their own search area within an entry or be indicated in different microstructural positions, or they may lack cross-referencing).

This results in the fact that time-consuming queries in parallel and comparable corpora are often required to obtain information concerning potential variants. Lexicographic resources are needed which provide users with terms together with relevant variants and variant-related information. A range of requirements can now be identified for lexicographic coverage of term variation:

- A need for systematic coverage of non-diasystemic variation;
- A need for syntactic and semantic information concerning variants;
- A need for pragmatic information concerning text sources, genres, type of communication;
- A need for a clear and coherent link between domain terminology and domain ontology.

As pointed out in Giacomini (2017), the operational and cognitive difference between the tasks of technical writers and professional translators do not change the fact that the main function of the envisaged dictionary should be to make variants and information about variants available in the native language of its users. Moreover, a lexicographic entry should consist of separate modules dedicated to the treatment of different information types. Each module should be separately accessible and information types should be combinable in order to enable users to perform targeted queries.

From a general structural perspective, a non-form-determined (conceptual) macrostructure and a form-determined macrostructure should be best combined (for classification of different types of macrostructures in electronic lexicography cf. Giacomini 2015). As a consequence, external data access should be made possible via both the ontological and frame-based path and the terminological path. In the proposed model, multiword terms and their variants appear both as a lemma and as another possible microstructural item (e.g. a variant or part of a corpus example) with related cross-references. Cross-referencing ensures a coherent representation of the different roles a term may play within the dictionary structure and, at the same time, reflects the relations existing between the different layers of the data model.

Given a multiword lemma as a main term, the abstract microstructure for a multiword term entry can be defined as follows:

ABSTRACT MICROSTRUCTURE
ontology-related data
frame-related data
language
lemma (main term, MT)
– syntactic and semantic structure
– corpus example(s)
– source(s)
– image
– variant
— syntactic and semantic structure
— variation type (O-, O+, M-, M~, M+, S-, S+)[4]
— corpus example(s)
— source(s)

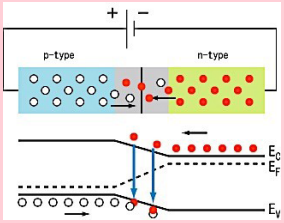
Further items may apply to specific microstructural data, but will not be the object of this paper.

The concrete microstructure related to the lexicographic entry of the multiword term *diode in forward bias* can be visualized as the composition of three descriptive areas corresponding to the three layers of data analysis, i.e. the terminology layer, the frame layer and the ontology layer. All information concerning relevant frame elements and ontological classes refers to both a term and its variants. An image[5] is also integrated into the lexicographic entry and is attributed to the main term.

4 The - and + symbols indicate absence or presence of a certain trait, while the ~ symbol, which is only contemplated in the case of morphology, stands for partial morphological variation.

5 S-kei CC-BY-SA-2.5 https://commons.wikimedia.org/wiki/File%3APnJunction-Diode-ForwardBias.PNG

TERMINOLOGY: EN	FRAME: Functionality	ONTOLOGY
MT: diode in forward bias $N_{Prod}\,p_{in}N_{Prop}$ The small signal model of a **diode in forward bias** is a resistance in parallel with a capacitance [inst.eecs.berkeley.edu] forward biased diode $V_{Prop}\,N_{Prod}$ O- M- S+ In a **forward biased** p-n junction **diode**, the positive terminal of the battery is connected to the p-type semiconductor material and the negative terminal of the … [physics-and-radio-electronics.com] **diode under forward bias** $N_{Prod}\,p_{under}N_{Prop}$ O- M- S- Figure 4.4.5: Current-Voltage characteristics of a silicon **diode under forward bias** [ecee.colorado.edu] **diode in forward direction** $N_{Prod}\,p_{in}N_{Prop}$ O- M~ S- First produced by Clarence Zener in 1934. It is similar to normal **diode in forward direction**, it also allows current in reverse direction when the applied voltage reaches the breakdown voltage. [Electronicshub.org]	- Product: **diode** - Property: **forward bias**	**SEMICONDUCTOR DIODE** - MATERIAL --- SEMICONDUCTOR MATERIAL --- HOUSING MATERIAL --- **PHYSICAL PROPERTY** - FORM --- COMPONENT --- CONSTRUCTION FORM --- HOUSING TYPE - FUNCTION --- MOUNTING TECHNOLOGY --- MOUNTING TECHNIQUE --- APPLICATION --- USER



Separate or combined queries involve each data type (i.e. microstructural item) available in the dictionary database. For instance, the output of a search query can be

(i) all variants of a multiword term,
(ii) specific orthographic, morphological or syntactic variants of a multiword term (e.g. O- M- S-),
(iii) multiword terms corresponding to a given syntactic structure with given POS content (e.g. N pN),
(iv) multiword terms matching a specific frame element or frame element combination (e.g. PRODUCT + PROPERTY),
(v) multiword terms matching a specific ontological class or class combination (e.g. terms matching the class PHYSICAL PROPERTY), or
(vi) frame elements and ontological classes matching a multiword term.

5 Conclusion and Further Work

This paper has introduced a frame-based description model for technical terms and their variants in an e-dictionary covering terminology related to the field of semiconductor diodes. The main goal of the paper was to provide information regarding methodology involved in developing the three layers of term and variant analysis (ontology – frame – terminology) and to introduce microstructural properties of the technical dictionary. Synonymous variation, especially in multiword terms, is at the center of discussion as a significant but still underestimated phenomenon in terminology. Electronic lexicography is a privileged area in which resources can be created to provide extensive coverage of this phenomenon. In this same area, an important contribution to the improvement of NLP procedures for term and variant extraction from specialized corpora can be made by exploring target users' needs and designing corresponding data models.

The applicability of the proposed approach to other technical domains is presently being tested on corpora containing texts about technical artefacts (thermal insulation products and DIY-tools) but referring to technical domains with different conceptualization, standardization and communicative features. Frame-based annotation of terms and variants turns out to be feasible provided that requirements for a coherent ontology and an exhaustive frame description with the right granularity (i.e. an appropriate level of semantic detail to ensure reliable annotation) are satisfied. Promising results obtained in the context of this paper as well as in the underlying project for what concerns frame-based data modelling and related corpus annotation lay a sound basis for future efforts towards a better lexicographic and terminographic coverage of multiword term variation in specialized language.

References

Caro Cedillo, A. (2004). *Fachsprachliche Kollokationen*. Tübingen: Gunter Narr.

Clouden, L. (2014). *Physics: A Concise Revision Course for CXC*. Cheltenham: Stanley Thornes.

Daille, B. (2005). Variations and application-oriented terminology engineering. Terminology 11.1.

Daille, B. (2017). *Term Variation in Specialised Corpora*. Amsterdam: John Benjamins.

Diffenderfer, R. (2005). *Electronic Devices: Systems and Applications*. Clifton Park: Thomson.

Faber, P. (2015). Frames as a framework for terminology. In: Kockaert, H.J./ Steurs, F. (eds.), *Handbook of terminology* (Vol. 1). Amsterdam: John Benjamins, pp. 14-33.

Faber, P. (ed.) (2012). *A cognitive linguistics view of terminology and specialized language*. Berlin/Boston: De Gruyter Mouton.

Fernández-Silva, S. & Kerremans, K. (2011). Terminological variation in source texts and translations: A pilot study. In: *Meta: Journal des traducteurs. Meta: Translators' Journal*, 56(2).

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

Freixa, J. (2005). Variación terminológica:¿ por qué y para qué?. In: *Meta: Journal des traducteurs. Meta: Translators' Journal*, 50(4).

Giacomini, L. (2017). An ontology-terminology model for designing technical e-dictionaries: formalisation and presentation of variational data. In *Proceedings of eLex 2017*, September 2017, Leiden (NL).

Giacomini, L. (2015). Macrostructural properties and access structures in LSP e-dictionaries for translation: the technical domain. In *Lexicographica* 31.2015, pp. 90-117.

Godse, A.P. & Bakshi, U.A. (2010). *Electronic Devices and Circuits I*, Pune: Technical Publications.

Jacquemin, C. & Tzoukermann, E. (1999). NLP for term variant extraction: synergy between morphology, lexicon, and syntax. In: *Natural language information retrieval*. Dordrecht: Kluwer Academy Publishers, pp. 25-74.

Reimerink, A. & Faber, P. (2009). Ecolexicon: A frame-based knowledge base for the environment. In *Proceedings of the International Conference Towards eEnvironment*, pp. 25-27.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C.R. & Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.

Salivahanan, S., Suresh Kumar, N. & Vallavaraj, A. (1998). *Electronic Devices and Circuits*. New Delhi: Tata McGraw-Hill.

Schmidt, T. (2014). *The Kicktionary: Combining corpus linguistics and lexical semantics for a multilingual football dictionary*. IDS Mannheim.

Temmerman, R. (2000). *Towards new ways of terminology description: The sociocognitive-approach* (Vol. 3). Amsterdam: John Benjamins.