The Sublanguage Factor

Modeling Term Variation in Clinical Records

Proefschrift aangeboden door

Leonie Grön

tot het behalen van de graad van

Doctor in de Taalkunde

KU Leuven
Faculteit Letteren

Promotor: Ann Bertels
Co-promotor: Kris Heylen
Openbare verdediging: 16 september 2019

Contents

List of Abbreviations

ADE	Adverse Drug Event
BCFI	Belgisch Centrum voor Farmacotherapeutische Informatie
CHV	Consumer Health Vocabulary
CITIP	Centre for IT and IP Law
CLEF	Conference and Labs of the Evaluation Forum
CRF	Conditional Random Fields
cTAKES	Clinical Text Analysis and Text Extraction System
CTT	Communicative Theory of Terminology
ED	Emergency Department
EHR	Electronic Health Record
EMC	Erasmus University Medical Center
FBT	Frame-Based Terminology
FSN	Fully Specified Name
GTT	General Theory of Terminology
HITECH	Health Information Technology for Economic and Clinical Health
HYPHEN	Hybrid System for the Normalization of Phenotype Concepts
IAA	Inter-Annotator Agreement

ICD	International Statistical Classification of Diseases and Related Health Problems
ICU	Intensive Care Unit
IHTSDO	International Health Terminology Standards Development Organisation
ISO	International Organization for Standardization
LCS	Longest Common Substring
LIIR	Language Intelligence & Information Retrieval
LSP	Linguistic String Project
MARS	Machine Reading of Patient recordS
MedLEE	Medical Language Extraction and Encoding System
MiPACQ	Multi-Source Integrated Platform for Answering Clinical Questions
ML	Machine Learning
MWE	Multi-Word Expression
NER	Named Entity Recognition
Nictiz	National ICT Instituut in de Zorg
NLP	Natural Language Processing
PHI	Personal Health Information
PoS	Part of Speech
PT	Preferred Term
QLVL	Quantitative Lexicology and Variational Linguistics

RFC	Random Forest Classifier
SCAN	Swedish Clinical Abbreviation Normalizer
SCTID	SNOMED CT Identifier
SEAM	Semi-Automated Ontology Management
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
UMLS	Unified Medical Language System
UZ	Universitair Ziekenhuis
WSD	Word Sense Disambiguation

List of Tables

List of Figures

Acknowledgments

A much-cited African proverb says that it takes a village to raise a child. Academic wisdom has it that, to write a dissertation, it is not enough to toil away in solitude; you will need a number of people to help you through your hardship. This is the point where I, finally, get a chance to express my gratitude to those who supported me over the past years.

First and foremost, I would like to thank my doctoral supervisor, Ann Bertels. Throughout the years, Ann coordinated the practicalities of my dissertation, revised numerous papers and chapters (twice!), and did so much more. Most importantly, she always encouraged me and gave me the freedom to pursue my ideas. I would also like to thank my co-supervisor, Kris Heylen, for his feedback and support with the practical organization of the research project. I am grateful as well to the members of my supervisory committee, Frieda Steurs and Maribel Tercedor Sánchez, who followed up on my progress and provided helpful feedback over the years. Also, I would like to thank the additional jury members, Sophia Ananiadou, Geert Brône and Sien Moens, for taking on the responsibility of serving on my examination committee.

My research would not have been possible without the collaboration of UZ Leuven. In particular, my thanks go to Chantal Mathieu and Carolien Moyson from the department of endocrinology, and Erwin Bellon and Kevin Renaers from ICT, for their help with the collection and annotation of the data.

Two people deserve special thanks for helping me with the final steps: Johanna Grön, who designed the cover of this book; and Markus Arntzen, whose proofreading skills made an invaluable contribution.

I would also like to thank all colleagues, past and present, at QLVL and elsewhere, for their help with daily matters and good company over the years.

Besides, I would like to mention Sylvia Lescrauwaet and Sonja Dekoning, who promptly assisted with all matters administrative.

In trying times, we turn towards our family. I am no exception. Thank you for being there and believing in me. Karlien, Kristina and Theresa: Thank you for your friendship.

Finally, at the very bottom, where they belong, I would like to express my gratitude to those people who seek joy and liberty in the underworld: my fellow cavers. Over the past years, we shared many moments of joy and misery in a wonderful world where research does not matter. Thank you!

Introduction

The shift towards electronic forms of medical documentation, in particular the mass adoption of the electronic health record (EHR), has led to the accumulation of huge collections of health-related data in digital format. The availability of such data opens opportunities for scientific innovations in numerous disciplines: Clinical data can, for instance, serve as the basis for epidemiological surveillance to anticipate pandemic outbreaks of a disease; it can enable the identification of adverse events, such as side effects of a pharmaceutical product, or interactions of one drug with another; and it can be used to evaluate the effectiveness of a treatment. In brief, there are many ways in which the reuse of health data could pave the way for advancing clinical research, improving health care systems, and, finally, ensuring the quality and efficiency of care provided to the individual patient. However, these opportunities also come with a range of methodological challenges: Which infrastructure is required to enable the electronic documentation of health care in a safe and consistent manner? Which ethical standards must be set to safeguard the personal rights of the patients, and how can these standards be translated into legal frameworks? And, essentially, how can the information contained in an EHR be translated into a form that can serve as input for scientific applications?

This thesis originated in the interdisciplinary project MARS (Machine Reading of Patient Records), which addressed some of these questions. The primary goal of this project was to bring together expertise from different fields to develop advanced methods for clinical data processing in the Belgian context. The project involved partners from four disciplines, namely the Universitair Ziekenhuis (University Hospital; UZ) Leuven, the Centre for IT and IP Law (CITIP), the Language Intelligence & Information Retrieval (LIIR) lab and the research group Quantitative Lexicology and Variational Linguistics (QLVL), where the research presented in this thesis was carried out. Methods from computational and variational linguistics, which are the focus of the research conducted at LIIR and QLVL, play a crucial role in

operationalizing health data for further use: As EHRs are largely composed in free text, natural language processing (NLP) is the key technology for unlocking medical facts from unstructured data. Moreover, the language used in EHRs deviates strongly from general written language, such that existing tools cannot be readily applied for the extraction and normalization of relevant information. The development of domain-specific NLP systems has thus become a vibrant area of research, which has brought forth a number of advanced applications for the processing of clinical text in English. For under-researched varieties like Belgian Dutch, though, resources are still scarce. This thesis contributes to closing this gap: Based on a case study involving the empirical analysis of a clinical dataset, it will analyze variation patterns in the usage of clinical terminology in Belgian Dutch. Such patterns can be leveraged to improve computational methods for the automatic processing of clinical language.

The thesis consists of three parts. Part I gives an overview of the context and theoretical framework: After outlining the potential benefits and applications of clinical data reuse, Chapter 1 describes the state-of-the-art of clinical NLP, and assesses the impact of term variation on its performance. Chapter 2 describes the major currents in terminological theory, from the traditional, strictly normative approach, to descriptive approaches inspired by socio-cognitive linguistics, which consider term variation as a subject of study in its own right. Based on the classification schemes developed by earlier work, a typology of clinical term variation is presented. Chapter 3 shifts the focus from the immediate lexical level to the wider linguistic context, and introduces sublanguage theory as a framework for the analysis of specialized languages. After describing the sublanguage properties of clinical writing, it gives an overview of earlier research employing sublanguage theory for the analysis and processing of this genre. Part II moves on to the empirical analysis, i.e. the terminological case study: Chapter 4 identifies gaps in previous research that motivate the case study, formulates its main hypothesis and gives an overview of its procedure. Chapter 5 introduces the EHR sample that is at the core of the analysis. Based on their stylistic and thematic properties, the individual sections of the EHRs are characterized as distinct sublanguages. To quantify term usage and variation in these sublanguages, the

dataset was annotated at two levels. Chapter 6 presents the first step, i.e. the manual annotation of the dataset with concept identifiers from a clinical terminology. The output of this annotation enables the characterization of the individual sublanguages by their conceptual and terminological structure, and the identification of semantic and pragmatic factors that determine the potential for variation. To further distinguish between individual types of variation, the terms were annotated at a second level, i.e. with formal features. Chapter 7 introduces the feature set used to characterize different term types, and quantifies their distribution across the annotated terms. The results demonstrate the interaction of conceptual properties with variation processes. In Part III, the insights gained from the descriptive analysis are validated by statistical means: Chapter 8 describes the final experiment, where term variation is modeled as a function of semantic features and context factors. The results illustrate the complex interactions of sublanguage properties with term variation, both at a global and a local level. To round off this thesis, Chapter 9 presents the final conclusion.

Part I

Chapter 1: The Impact of Term Variation and Sublanguage Features on Clinical Data Reuse

With the mass adoption of the EHR, a range of new opportunities for clinical data analysis has opened up; the development of NLP techniques suitable for the domain has since become a priority. Handling term variation remains one of the major challenges in the field.

The first two sections of this chapter outline the circumstances that drove the rise of the EHR as the standard medium for health documentation (Section 1.1), and give an overview of the major areas of clinical research that can benefit from clinical data reuse (Section 1.2). The third section illustrates how NLP methods are leveraged for the automatic processing of clinical text, and how these methods are affected by term variation (Section 1.3). The final section summarizes opportunities and challenges for clinical NLP, in particular with regard to low-resource languages like Belgian Dutch (Section 1.4).

1.1 The Rise of the EHR

In the first decade of the 21st century, the U.S. government passed a number of laws that would fundamentally transform health documentation: In 2003, the electronic submission of codes for diagnoses and procedures became a requirement for reimbursement within the national health care programs (Meystre et al. 2017). In 2009, the Health Information Technology for Economic and Clinical Health (HITECH) Act was passed, which encourages the use of electronic health documentation to improve the quality and efficiency of care. Crucially, HITECH foresees a remuneration for health care providers who fulfill certain structural criteria, and who implement an electronic form of documentation for "meaningful use" (Adler-Milstein and Jha 2017, 1416).

This act thus provides a strong financial incentive to switch to the EHR, resulting in a substantial increase of EHR adoption rates among eligible institutions. In other countries as well, including Australia, Canada, Denmark, Finland, and the U.K., political measures were taken to stimulate the use of electronic services for health care (Gardner 2016; Kaipio et al. 2017). For instance, in Portugal, the national health service launched an EHR portal that allows patients to access their reports, schedule appointments and request drug prescriptions (Tavares and Oliveira 2017). In Belgium, a federal network for the secure storage and exchange of health data was set up in 2008 (France 2011); in subsequent years, the ministry of health passed a series of action plans in order to support the implementation of electronic services among providers, but also to improve health literacy among patients (De Block et al. 2019).

In the course of the past two decades, the EHR has found its way into medical practice. While there are still major differences in usage rates, depending, for instance, on structural aspects of the national health system and the age group of the users (Evans 2016), the EHR is on its way to become the standard form of health documentation.

1.2 Fields of Clinical Data Reuse

While the public debates around the EHR and its legal and ethical implications have only recently gained momentum, the medium itself is not such a new invention: In fact, efforts for the digital collection of health data go back to the 1970s (Meystre et al. 2017). However, the EHR has long stayed confined to the academic setting, as most health care providers lacked the necessary ICT infrastructure and technical expertise to use it. Besides, the EHR had no clear advantage over the traditional paper record in daily care; there was thus no good reason to change the running system. With the advent of affordable computers and user-friendly software, though, the benefits of digital documentation became more obvious (Evans 2016). Initially, many physicians appreciated the EHR mainly for alleviating the burden of administration; however, clinical research soon began to explore possibilities of

using the collected data for additional tasks, such as decision support (Gardner 2016). The collection of more comprehensive health databases since the 2000s coincided with a rising interest in data science, both for scientific and commercial purposes. The benefit of clinical data reuse, i.e. the "non-direct care use of personal health information including but not limited to analysis, research, quality/safety measurement, public health, […] and marketing and other business including strictly commercial activities" (Safran et al. 2007, 2) is now widely accepted (Martin-Sanchez and Verspoor 2014).

The secondary use of clinical data has thus become a most vibrant area of research. Clinical data serves both as the basis for *information extraction*, i.e. the extraction and aggregation of known facts (e.g. for quantifying the prevalence of a disease among a population); and for *data mining* or *text mining*, i.e. the discovery of new knowledge from structured or unstructured data (e.g. for detecting associations between drugs and hitherto unknown side effects; cf. Ananiadou and McNaught (2006) and Meystre et al. (2008) for the distinction).

Given the constant increase in the rates of related publications, it is increasingly difficult to keep up with new developments in the field. While the potential applications of clinical data reuse are manifold, a recent review by Wang (2018) identifies three major topics: *diseases*, *drugs* and the *clinical workflow*.

1.2.1 Diseases

For diseases, most studies focus on phenotyping, i.e. the identification of cases that meet a pre-defined set of symptomatic or diagnostic criteria. Phenotyping is thus a key method for building reliable patient cohorts, both for retro- and prospective analysis (Ford et al. 2016). For instance, the identification of respiratory tract symptoms in clinical notes allows the large-scale monitoring of influenza, which enables both the modeling of seasonal patterns (Chapman, Chu, and Dowling 2007), and the surveillance of acute outbreaks (Elkin et al. 2012). Apart from detecting the prevalence of a given

disorder, such as peripheral artery diseases (Savova et al. 2010a), phenotyping can also serve to capture more fine-grained properties. Carrell et al. (2014) model the probability of cancer recurrence for breast cancer patients; Skevofilakas et al. (2010) predict the onset of a secondary disease, namely retinopathy, which is a common complication among diabetes patients. Moreover, phenotyping can reveal co-morbidities that are not explicitly documented, such as psycho-social factors and clinical observations (Boytcheva 2012).

1.2.2 Drugs

Among the studies related to drugs, one central theme is pharmacovigilance, i.e. the monitoring of drug safety. A growing body of work is devoted to the detection of adverse drug events (ADEs). For instance, Wang et al. (2009) detect ADEs for seven drug classes, including widely prescribed substances such as ibuprofen. Besides, clinical data can provide input for pharmacogenetic studies. Xu et al. (2011a) propose a system that first identifies weekly doses of warfarin in a patient sample, and then associates the dosage with genetic variants. Apart from the effect of the pharmaceutical agent itself, the main cause of adverse events is incorrect administration. Clinical data can be used to detect both errors in prescription, and in drug uptake. For instance, Breydo, Chu, and Turchin (2008) present a method to detect inactive medications. They find that a substantial part of the EHRs in the analyzed sample (one in five documents) contains drugs that were discontinued at an earlier point, but had not been removed from the list of active medications. Another risk factor is poor compliance of the patient. Turchin et al. (2008) present an algorithm to identify cases of non-adherence, i.e. patients that refuse to take their medication or do so only sporadically. Carrell et al. (2015) study the opposite phenomenon: They detect overuse of prescription opioids among hospitalized patients.

1.2.3 Clinical Workflow

Clinical data is a key resource to improve the clinical workflow, both at the individual and the structural level. While the support of basic administrative tasks like billing is still an important task (e.g. Perotte et al. (2014)), the scope of applications has widened considerably. For instance, Bozkurt et al. (2016) extract detailed descriptors from mammography reports to inform a decision support system for early-stage cancer detection. Ruud et al. (2010) mine discharge letters for details of follow-up interventions. As the timely scheduling of follow-ups may prevent disease recurrence, this information can be used to reduce costly re-admissions to the hospital at a later stage. Another unnecessary cost factor are inappropriate admsissions, especially visits to the emergency department (ED). By analyzing primary care notes, St-Maurice, Kuo, and Gooch (2013) identify concepts that are strongly correlated with inappropriate ED use, such as psycho-social and pain-related disorders. The immediate assessment of such conditions could thus help to efficiently reduce the ED workload.

1.3 The Role of NLP in Clinical Data Reuse

The reuse of clinical data requires that the relevant information be represented in a way that is suitable for systematic analysis. However, while most EHR systems provide a template with separate fields to capture particular types of information, most of these fields do not have a specific input format. Apart from findings expressed in numerical format, such as laboratory results or measurements, clinical documentation is thus mostly done in free text: In a survey among U.S. hospitals, Cannon and Lucci (2010) found that almost two thirds (65%) of the EHR data was unstructured. Since the early days of electronic documentation, a number of projects pushed for an increase in structured data entry, for instance to support logical inferences in decision support (e.g. Litzelman et al. (1993)). However, a recent review found that the lack of normalized, interoperable formats is still one of the major barriers for the efficient reuse of health data (Kennell, Willig, and Cimino 2018). The main reason for this discrepancy is that free text remains the medium of

choice among clinicians: It is considered more intuitive, efficient and expressive than structured formats, such as concept codes from a controlled terminology (Bansler et al. 2016; Groth Jensen and Bossen 2016; Kaufman et al. 2016; Rosenbloom et al. 2011). Besides, it allows for the formulation of semantic nuances, such as intermediate categories or preliminary findings, which may lack an adequate representation in standardized coding systems (Ford et al. 2016).

NLP is thus the key technology for identifying relevant information in text and mapping it to a format that enables semantic interpretation (Velupillai et al. 2015). However, most NLP tools developed for general language perform poorly on this task. EHRs are composed in an environment where efficiency is imperative; therefore, they abound with shorthand expressions and simplified constructions that defy the conventions of general grammar. Moreover, as they are mostly intended for communication among peers, they contain a highly specialized vocabulary. The language found in EHRs has thus been characterized as a distinct *sublanguage* (cf. Chapter 3 for a detailed discussion). To handle this sublanguage, the development of domain-specific methods is required. Clinical NLP, i.e. "natural language processing methods developed and applied to support health care by operationalizing clinical information contained in clinical narrative" (Demner-Fushman and Elhadad 2016, 224), has thus emerged as a separate discipline.

The remainder of this section outlines the role of NLP in the secondary use of health data. While the greatest part of research has focused on the English language, there is growing interest in the automatic processing of EHRs in smaller languages. Even though the tasks themselves overlap, the lack of resources makes clinical NLP in such languages an even more demanding task. Therefore, the two are discussed in separate sections: First, Section 1.3.1 gives an overview of general methodological trends in different areas of clinical NLP; then, Section 1.3.2 sketches the state of the art in clinical NLP in languages other than English. Both subsections discuss three main areas: Firstly, research on the creation of *clinical corpora* for the development and evaluation of NLP applications; secondly, studies on the development and maintenance of structured *knowledge sources*, which are an important re-

source for many NLP systems; thirdly, work on *foundational methods* for the processing of clinical text, covering the specific tasks that arise at the different levels of an NLP pipeline.

1.3.1 Areas of Clinical NLP

1.3.1.1 Clinical Corpora

The availability of domain-specific corpora is indispensable for the development and evaluation of innovative applications, such as recognizers for medical entities. However, in the clinical domain, there is one major obstacle to the distribution of such datasets, namely *privacy* laws. While the exact legislation varies, most countries require the informed consent of the patient before any information on their case can be used for secondary purposes; if this consent cannot be obtained for practical reasons, the data must be de-identified. Concretely, this involves the removal or replacement of protected health information (PHI), such as names and identifiers used by the insurance system (Meystre et al. 2017). Certain types of PHI can be replaced by relatively simple methods: For instance, administrative identifiers, such as social security numbers, can be captured by regular expressions; personal and geographical names can be identified by dictionary lookup (e.g. Neamatullah et al. 2008). However, the blind scrubbing of all names and places might result in the loss of medically relevant information, such as eponyms and geographical names referring to the center of a disease outbreak. Meystre et al. (2014) compare the informativeness of EHRs before and after automatic de-identification. They conclude that the overall difference is small, but considerable. An additional challenge lies in the presence of variants, such as abbreviations of locations, which might be missed by purely knowledge-based methods. The development of reliable systems for de-identification thus remains a high priority (Kushida et al. 2012; Meystre 2015).

The training and testing of methods for clinical NLP further requires the availability of a reference standard annotated with semantic labels, such as concept codes from a medical terminology, and linguistic information, such as Part of Speech (PoS). As the manual *annotation* by human experts is

expensive, both with regard to temporal and financial resources, various strategies have been proposed to keep the costs at bay (Velupillai et al. 2015): For instance, statistical measures can be utilized to determine the number of documents required for the creation a representative reference standard (Juckett 2012). Automatic pre-annotations can also help to reduce the amount of manual labor and improve the consistency between human annotators (Grouin and Névéol 2014). Moreover, the required efforts depend crucially on the complexity of the annotation scheme. Therefore, in many projects, the scope is limited to the information required for a specific task, which is, however, coded in a fine-grained manner: For instance, Iqbal et al. (2017) present a pipeline for the detection of adverse events caused by antipsychotics. For development, they annotate a corpus of psychiatric records with side effects that are specific to this type of medication. On the other hand, higher-level linguistic properties, such as syntactic functions or semantic roles have a lower priority for most applications. However, a number of recent initiatives, especially in the context of shared tasks, have fostered the development of more comprehensively annotated datasets (Savova et al. 2017). One such example is the Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ) clinical corpus (Albright et al. 2013). This corpus was fully annotated with word features and syntactic structure at the formal level, and a subset of concept codes from a domain ontology at the semantic level. By adapting established annotation standards, the interoperability with existing resources was ensured. The development of such richly annotated corpora, and their dissemination among the research community, is an essential contribution to the development of generalizable methods, especially those involving deeper linguistic processing.

1.3.1.2 Knowledge Sources

Apart from domain corpora, the second pillar of clinical NLP are structured knowledge sources. Domain terminologies and ontologies provide a standardized semantic framework, which is required for operationalizing textual information. Besides, they are key resources for text-processing methods themselves: They provide curated lists of terms and semantic properties, which serve as input for both knowledge-based and machine-learning (ML) methods.

On the contrary, in clinical practice, the primary purpose of terminologies is to label EHRs with a set of unique codes, either for administration, or for the documentation of care. The most commonly used terminologies are the International Statistical Classification of Diseases and Related Health Problems (ICD) and the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT). While the ICD is considered easier to use by practitioners, SNOMED CT is more expressive (Dalianis 2018). In SNOMED CT, each concept is represented by at least one standard term and a set of semantic attributes. In addition, a concept entry may contain one or more established synonyms.

Increasingly, EHRs themselves are leveraged for the expansion and maintenance of existing terminologies. Clinical writing is a valuable resource for term acquisition, especially for that of lay and non-standard forms. For example, Henriksson et al. (2013) use distributional semantics to identify synonyms of SNOMED CT terms. Chen et al. (2017) present a hybrid system to extract and rank term candidates from a sample of discharge notes. The goal of this application is to prioritize terms for inclusion in a structured terminology, the Consumer Health Vocabulary (CHV; Zeng and Tse 2006). In addition, clinical text can be exploited to infer conceptual attributes and semantic relations. Pattern-based methods, as first proposed by Hearst (1992), are still a popular approach: For example, the Semi-automated Ontology Management (SEAM) system (Doing-Harris, Livnat, and Meystre 2015) relies on a set of lexico-semantic patterns to extract hierarchic relations from biomedical and clinical writing.

1.3.1.3 Foundational Methods

The particular features of the clinical sublanguage affect all levels of automatic processing, from basic pre-processing tasks such as segmentation, over the processing of individual words and the analysis of syntactic structure, up to the semantic interpretation of larger text portions.

Clinical text makes unconventional use of punctuation, which can complicate *word segmentation*. Like other genres of informal writing, some parts of an EHR can lack punctuation altogether. On the other hand, orthographic mark-

ers are frequently used for the formation of compounds or abbreviations. Particular character sequences can even take on a symbolic meaning in context (e.g. '++' to denote the progressive evolution of a symptom). Obviously, such expressions should be interpreted as a single token and not be discarded or split up. A similar problem occurs with the scientific notation of medical measurements, which may involve alphanumeric characters as well as punctuation marks (e.g. 'BP 140/90' to specify a blood pressure measurement). Nguyen and Patrick (2016) employ a pattern-based method to recognize different types of measurements in text and normalize them without introducing additional boundaries.

For processing at the word level, such as *PoS tagging,* the major challenge lies in the high proportion of domain-specific terms, including non-canonical forms. Clinical text not only contains standard medical terminology, but is also ripe with non-canonical variants. In addition, the PoS distribution in clinical text differs from that in general language; the probabilistic tagging of unknown words is thus error-prone (Pakhomov, Coden, and Chute 2006). As a result, off-the-shelf taggers often perform poorly on the clinical genre (Ferraro et al. 2013). The inclusion of lists of domain-specific terms drawn from medical knowledge sources can improve tagging accuracy to some extent; however, it will not solve the case for non-standard variants of these terms. Besides, the re-training on an annotated reference standard can support the domain adaptation of existing tools (Fan et al. 2011; Knoll et al. 2016); but, of course, this requires that such a dataset be available in the first place. For a thorough handling of non-dictionary words, including misspellings, abbreviations and ad-hoc word formations, customized strategies are required.

As EHRs are mostly not intended for public display, orthography is a minor concern. Misspelled words can thus make up a substantial portion of the vocabulary. While the exact numbers vary across document types and clinical specialties, Dalianis (2018) quotes figures ranging from 1.1% up to 11%. The normalization of misspellings can thus substantially improve the outcome of the subsequent steps in an NLP pipeline. Lai et al. (2018) present a comprehensive system for *spelling correction* in clinical text: The algorithm first

identifies misspellings based on dictionary lookup, then uses edit distance to suggest correction candidates, and finally ranks them according to the noisy channel model. One drawback of this method is that the ranking does not take the lexical context into account; this increases the chances that a lexically similar, but contextually inappropriate candidate will be selected. To tackle this problem, Fivez, Šuster, and Daelemans (2017) develop a model that includes contextual cues to determine the best correction: To rank the candidates, they compute the similarity between the candidate vector, and the context vectors of the misspelled form. Workman et al. (2019) combine distributional information with a string similarity metric and statistical information. For evaluation, they use two data samples, including pathology reports and ED notes. In a quantitative analysis, they note that, while the types of misspellings are similar, their distribution differs across the clinical specialties.

Similarly, as brevity is a driving principle of clinical documentation, EHRs tend to contain a high proportion of shorthand notations, many of which are ambiguous. Depending on the clinical specialty, the proportion of abbreviated forms can account for up to 14% of the medical terms (Dalianis 2018). Various strategies have been proposed for *abbreviation expansion* and *disambiguation*: The easiest way is to look up the long form in an abbreviation dictionary, such as the one created by Moon et al. (2014). However, this approach does not allow for the disambiguation of word senses in context. Moreover, as abbreviations tend to be created on the fly, their surface forms show a high degree of variation; the compilation of an exhaustive dictionary is thus not realistic. Therefore, the popularity of data-driven methods is increasing: Finley et al. (2016) use word vectors to disambiguate a large set of general medical abbreviations in clinical text. Their approach yields stable results, even on heterogeneous data sets. Liu et al. (2015) use word embeddings derived from biomedical writing and online sources to disambiguate abbreviations in intensive care unit (ICU) notes.

The syntactic structure of clinical text can deviate strongly from grammatical conventions. To speed up their writing, clinicians tend to use telegraphic constructions, omit function words or verbs. Due to these extra-grammatical

features, the *parsing* of clinical text is a challenge for state-of-the-art parsers, which are usually trained on annotated data from the general domain. Fan et al. (2013) extend existing annotation guidelines to cover ill-formed sentences and re-train an existing parser on a clinical corpus. However, as Kate (2012) points out, even within the clinical domain, there are major differences in terminology and stylistic properties. To ensure robust performance, a parser would need to contain one model per clinical specialty and document type. Given the high investment required for the compilation of training data, he proposes to use unsupervised techniques for the automatic induction of sublanguage grammars. The proposed system uses text tagged with PoS and semantic classes as input, and generates parse trees based on the calculation of the production costs. In the end, though, full parsing is a rather costly process, which might not even benefit the end application (Jiang et al. 2015; Yan Wang et al. 2015). Therefore, many systems only carry out a shallow syntactic analysis, using rules or semantic patterns to chunk the text and obtain the units of interest.

Ultimately, the goal of most systems is to translate the relevant information to a semantic schema that is suitable for machine analysis. In the clinical domain, *Named Entity Recognition (NER)* typically involves the identification of medical entities and their subsequent mapping to a structured knowledge source. A number of ready-made processing suites exist for the task: MetaMap identifies entities from free text to concepts in the Unified Medical Language System (UMLS; cf. Bodenreider (2004)). The clinical Text Analysis and Text Extraction System (cTAKES; cf. Savova et al. (2010b)) maps entities to the corresponding identifier in SNOMED CT. However, most of the commonly used systems draw primarily on dictionary lookup to identify relevant entities; the recognition rate thus strongly depends on the coverage of the term base. One strategy to improve the mapping of non-canonical terms is to couple a knowledge-based recognizer with a module for variant generation. For instance, Thompson and Ananiadou (2018) present a hybrid system for the normalization of phenotype concepts (HYPHEN). The system employs an ensemble of rule-based strategies to generate variants for dictionary terms and normalize the forms encountered in text; in the evaluation, it outperformed MetaMap on both biomedical and

clinical data. Alternatively, it can be beneficial to enhance the built-in dictionary with a domain lexicon, especially for the processing of highly specialized genres such as radiology reports (Hassanpour and Langlotz 2016). Data-driven methods, on the other hand, do not suffer from the limitation of term coverage. For instance, Jonnalagadda et al. (2012) use distributional semantics to identify medical entities. As the recognition method is agnostic to the surface form of the entity, but relies only on its context, it is more sensitive to entities expressed in non-standard terms. Still, most state-of-the-art systems have a hybrid architecture, combining ML with rules and features derived from external knowledge sources (Pradhan et al. 2015).

The full semantic interpretation further requires the analysis of contextual properties, such as the factuality and temporal sequence of events. *Negation detection* is a well-studied task. The most popular algorithms, such as NegEx (Chapman et al. 2001), rely on trigger rules and lexical cues to identify cases of uncertainty or negation. Conversely, Wu et al. (2014) develop a ML approach and evaluate it on different types of EHRs. They report that, while the system can be fine-tuned to excel on a particular dataset, it is difficult to generalize the model for robust performance on a mixed sample. With regard to the *extraction of temporal relations*, there are still many remaining challenges (Kreimeyer et al. 2017): As Sun, Rumshisky, and Uzuner (2013) point out, many temporal expressions encountered in text are intrinsically ambiguous, as they are defined in relation to other points in time, or because their duration can only be inferred by world knowledge. Consequently, few systems produce reliable results, especially when presented with a heterogeneous dataset (Bethard et al. 2018).

1.3.2 Clinical NLP in Languages Other than English

While healthcare is a global priority, there is a huge discrepancy between the state of the art of clinical NLP in English and in other languages. Evidently, no other language is backed up by such a strong and well-connected research community as English: In a recent survey of the field, Névéol et al. (2018a) report that only 10% of the publications on clinical NLP focus on another

language. On the whole, though, there are encouraging signs that the discipline is gradually opening up towards other languages (Névéol and Zweigenbaum 2018).

1.3.2.1 Clinical Corpora

One of the main obstacles to methodological advances in non-English languages is the lack of corpora with comprehensive annotations. Compared to English, where the creation and dissemination of annotated datasets was fostered by numerous shared tasks, few such challenges have been organized for other languages. Among the exceptions are the most recent editions of the CLEF (Conference and Labs of the Evaluation Forum) eHealth Evaluation Lab (Goeuriot et al. 2017; Névéol et al. 2016, 2018b), which led to the creation of shareable datasets in other languages, including French, Hungarian and Italian. Likewise, clinical datasets in Japanese datasets have been made available in the context of the MedNLP tasks (Aramaki et al. 2014, 2016). Another noteworthy project is the Turku Clinical Corpus, which consists of nursing notes in Finnish. These notes were fully annotated with syntactic structure and argument roles, resulting in a full Treebank and PropBank that are freely available online (Haverinen et al. 2010, 2015; Laippala et al. 2014).

Apart from that, though, most datasets are strongly tied to individual research groups and difficult to access (Dalianis 2018). For example, while previous work on clinical German has studied 10 different corpora, none of them is publicly available (Lohr, Buechel, and Hahn 2018). In addition, as most of them were developed in the context of a particular project, they are only annotated with features that were relevant for the task at hand. For instance, Grouin and Névéol (2014) compile a mixed French corpus, containing clinical notes from different hospitals and specialties. This corpus was annotated with different types of PHI to serve as reference for automatic de-identification. Afzal et al. (2014) create a Dutch clinical corpus, comprising various document types, such as radiology reports and discharge summaries, from different institutions. To train a tool for the disambiguation of contextual properties, a subset of medical entities was coded for temporality, factuality and experiencer. He et al. (2017) compile one of the few non-English

corpora with comprehensive annotations: This dataset, which comprises discharge letters and clinical notes from a Chinese hospital, was manually coded with syntactic features and semantic properties, resulting in a rich resource for methodological developments.

Next to the creation of monolingual corpora, some attempts have been made to combine and align data from different languages. Such multilingual datasets enable cross-linguistic and -cultural comparisons, which can provide valuable insights about the portability of NLP methods. For instance, Allvin et al. (2011) study the linguistic properties of nursing notes in Finnish and Swedish. They observe strong similarities with regard to the document structure, the vocabulary and linguistic alternations, even though the two languages are unrelated. Wu et al. (2013) analyze two samples of clinical notes composed at U.S. and Chinese institutions. Here, the differences are more evident, in particular with regard to the proportion of semantic classes among the medical entities.

Given the lack of training data, and national differences in the notation of personal details such as street addresses, the *de-identification* of clinical data is mostly tackled by knowledge-based approaches. For example, for the de-identification of German admission notes, Richter-Pechanski, Riezler, and Dieterich (2018) enhance a general-domain recognizer with gazetteers, rules and regular expressions. In addition, to capture spelling alternatives, they implement an algorithm that generates lexical variants based on minimal edit distance. They note that, while the system performed well for regular types of PHI, such as post codes, which can be recognized in a purely rule-based manner, the de-identification of more ambiguous types, such as personal names, was still error-prone.

1.3.2.2 Knowledge Sources

With regard to terminology as well, the shortage of validated resources is a key issue. While a number of knowledge sources is available in other languages than English, the conceptual and terminological coverage of the local extensions remains relatively poor: In the 2016 release of the UMLS, the number of terms available in the subsets for other languages, such as Dutch,

was less than 5% of the number of terms in English (Névéol et al. 2018a). Similarly, for SNOMED CT, there are a number of local extensions in the making; however, as of now, full releases are only available for Argentine Spanish, Danish and Swedish. For other languages, such as Belgian Dutch, the current versions only cover a minor part of the original set of concepts (Mertens 2018). In addition, the translated versions only provide one standard term per concept, but no synonyms. Therefore, they are usually insufficient to serve as the sole basis for knowledge-based NLP applications (Skeppstedt et al. 2014).

A number of automatic strategies have been explored to expedite the translation process: For example, Deléger, Merkel, and Zweigenbaum (2009) compile a parallel corpus of biomedical articles in English and French. Using word alignments, they acquire translations candidates for SNOMED CT terms. Similarly, Heyman, Vulić, and Moens (2018) experiment with neural networks for the acquisition of English-Dutch translation pairs from a comparable corpus of medical Wikipedia articles. Their system uses both word and character-level representations for the classification of translation candidates. They note that, due to the high proportion of orthographically related forms in the medical domain, such as terms based on neoclassical roots, character-level representations are particularly useful. Schulz et al. (2013) and Cornet, Hill, and De Keizer (2017) assess the potential benefit of machine translation for porting SNOMED CT to German and Dutch respectively. While both studies report that off-the-shelf tools perform surprisingly well, the generated translations are not considered fit for the direct integration into controlled terminologies. Given the high quality standards, manual validation remains a mandatory and time-consuming step.

While fully-fledged terminologies are underway, the creation and expansion of local term bases is required to bridge the gap. A number of studies in different languages have exploited biomedical text, in particular journal articles, for the enrichment of existing vocabularies with standard medical terms (e.g. Drame, Diallo, and Mougin 2012; Skeppstedt and Henriksson 2013; Vagelatos et al. 2011). Clinical data, on the other hand, has proven a valuable source for the acquisition of terms that are specific to a particular

clinical specialty, as well as informal variants and abbreviations. For example, Bretschneider, Zillner, and Hammon (2013) design a semantic grammar to harvest domain-specific terms from a sample of radiology reports. By exploiting distributional properties, Henriksson et al. (2014) identify non-canonical synonyms and abbreviations in a mixed corpus of Swedish journal articles and clinical records. Xu et al. (2015) use term alignments to build a Chinese-English dictionary from two samples of discharge summaries. Starting from a seed set of terms from an existing terminology, they extract translation candidates based on contextual similarity. Zhang et al. (2017) present a hybrid method for enriching the Chinese SNOMED CT: For the recognition of term candidates, they use regular expressions and a classifier based on conditional random fields (CRFs); for the subsequent mapping to SNOMED CT concepts, they combine two metrics based on lexical and conceptual similarity.

1.3.2.3 Foundational Methods

For the greatest part, the sublanguage features of clinical writing are a product of the creation of their circumstances, not of a specific language. Therefore, the methodological challenges faced by clinical NLP in other languages largely overlap with those encountered in the processing of clinical English. This can enable the cross-lingual transfer of existing systems, especially for higher-level applications. However, for the core NLP tasks, customized methods are required to deal with the particularities of the language under investigation.

For languages where word boundaries are not marked explicitly in the orthography, *segmentation* requires increased attention. Cohen (2012) reports that, in the processing of clinical text in Hebrew, medical loanwords can be difficult to handle. As off-the-shelf tools rely on PoS features to detect word boundaries, unknown words cause segmentation errors. To address this issue, the segmentation tool is enhanced with a custom dictionary of transliterated medical terms.

As most languages have a richer morphological system than English, a more profound analysis might be required at this level. In many other Germanic

languages, nominals are inflected, which increases the number of surface forms encountered in text. As long as the declination patterns of the local language are observed, off-the-shelf tools can be used for *stemming* or *lemmatization*. However, many medical terms are derived from Greek or Latin, which come with their own inflectional system. Despite efforts for standardization, forms inflected according to the rules of the source language, and forms adapted to the native morphology, are used interchangeably in clinical records. In such cases, customized strategies for term normalization are needed (cf. Grigonyte et al. (2016) for Swedish; Spyns (1996) for Dutch).

Another process that can cause term proliferation in Germanic languages is compounding. As compound nouns allow for the concise expression of complex entities and events, their proportion in specialized texts, such as clinical writing, is particularly high: Bretschneider and Zillner (2015) report that, in a sample of clinical notes from a German hospital, the proportion of compounds among all nouns amounts to more than 60%. As many of these compounds are genuinely new formations, which are not covered by existing terminologies, *compound splitting* is an important step in an NLP pipeline. Spyns (1996, 2000) proposes a rule-based strategy to handle compound terms in clinical Dutch: Compounds are detected based on PoS sequences and split into their components by dictionary lookup. Bretschneider and Zillner (2015) use a semantically informed system for decompounding in German. Their algorithm relies on a domain corpus to identify compounds and generate splitting options, and then uses ontological relations and statistical measures to determine the best split.

As reduced forms are pervasive in clinical writing across languages, *abbreviation detection* and *expansion* is a well-studied problem. However, as the dominant reduction patterns vary, existing tools cannot be ported directly, but must be fine-tuned to the language under investigation, and, ideally, the clinical specialty. For example, in German, the period symbol is conventionally used to mark shortened forms; the task of abbreviation detection thus overlaps with that of sentence segmentation. To tackle this problem, Kreuzthaler and Schulz (2015) propose an ML method: Using a rich feature set, including the orthographic form of the token, its context, and dictionary

matching, they train a classifier based on support vector machines. Conversely, in a study of Hungarian ophthalmology report, Siklósi and Novák (2013) encounter the opposite problem: While, according to the rules of standard orthography, the period should be used to mark abbreviations, it tends to be omitted in their dataset. Therefore, they propose a rule-based algorithm which first detects abbreviations based on token features, such as consonant-vowel ratio and word length; then, for expanding the abbreviated forms, the system generates regular expressions and matches them against a domain dictionary and the corpus itself. Isenius, Kvist, and Velupillai (2012) present the Swedish Clinical Abbreviation Normalizer (SCAN), which employs rules, heuristics and knowledge sources to identify and expand abbreviations. While the system was originally developed for the processing of ED notes, Kvist and Velupillai (2014) present an updated version, which is also evaluated on radiology reports. They note that, even within one language, the proportion of abbreviation types varies across clinical specialties. Rubio-López, Costumero, and Ambit (2017) address the problem of abbreviation disambiguation for Spanish with an ML approach. A manually annotated dataset, consisting of clinical notes relating to stroke patients, is used to train a classifier on word- and document-level features. They find that, since the intended meaning of a form typically stays constant over a text passage, the section within the document is a particularly informative feature.

At the word level, the main challenge is identical with that encountered in English, namely that existing knowledge sources do not adequately reflect the vocabulary encountered in clinical writing. For the *PoS tagging* of clinical Dutch, Spyns (1996, 2000) enhances a knowledge-based tagger with morphological analysis: The tagger operates primarily by dictionary lookup; unknown words are passed on to a category guesser. As most of these are composed of a limited set of Greek or Latin confixes, this strategy achieves satisfactory results; however, the handling of neoclassical terms which are inflected according to the pattern of the source language remains an unresolved issue. In more recent work, the trend goes towards the domain adaptation of existing parsers: Wermter and Hahn (2004) evaluate the performance of two off-the-shelf taggers, one statistical and one rule-based, on a sample of clinical notes in German. They find that, even without training on domain

data, the statistical tagger performs surprisingly well on the task. They conclude that this result can be attributed to the properties of the clinical sublanguage, where the distribution of PoS sequences is rather limited. However, most subsequent works agrees that domain adaptation is crucial to achieve accurate results: For instance, Oleynik et al. (2010) compare the performance of a PoS tagger for Portuguese after training on general and domain data. They report that training on a curated gold standard of discharge summaries effects a substantial increase in accuracy. Hellrich et al. (2015) present a machine-learning tagger as part of a toolkit for processing clinical German. The tagger employs domain-specific token features, such as the number of Greek characters, and draws on an extended tagset to distinguish between sublanguage-specific categories, such as Latin terms in nominative or genitive case, which might influence further processing.

Similar to abbreviations, misspellings are a common feature of clinical text across all languages. *Spelling correction* is therefore an important step in term normalization. Dziadek, Henriksson, and Duneld (2017) evaluate the effect of different correction methods on the recognition of SNOMED CT concepts in Swedish biomedical and clinical text. For the ranking of correction candidates, they experiment with a number of different criteria, including string similarity, PoS values and context tokens. They find that the inclusion of context information produces the best results.

For the *parsing* of clinical text, Spyns (1996, 2000) employs a rule-based approach. Based on the analysis of an EHR sample in Belgian Dutch, he develops a sublanguage grammar. Drawing on this grammar, the parser is able to handle unconventional syntax as well as domain-specific semantic constellations, such as non-canonical verb valency patterns. However, given the extreme variability of clinical language, paired with the lack of formal constraints, the coverage of the parser remains limited. Obviously, a purely knowledge-based approach quickly reaches its limit. Therefore, more recent work has focused on the domain adaptation of language-independent statistical parsers. Laippala et al. (2014) compare the performance of two dependency parsers on different clinical sublanguages in Finnish. For domain adaptation, they use three treebanks, containing reports from intensive care and

cardiology, as well as daily nursing notes. The language used within the individual samples differs with regard to vocabulary size, as well as the average length and complexity of the sentences. Still, they find that the best results are achieved by including the entire dataset for training. He et al. (2017) re-train two dependency parsers on a manually annotated corpus of Chinese clinical notes. By combining features from both parsers, they achieve excellent results; however, they note that one clear limitation of their study is the homogeneity of the test data, which contains only two document types from a single institution. To overcome this limitation, Kara et al. (2018) create fictitious data for evaluation. To adapt a parser to clinical German, they first annotate a reference standard of nephrology reports with PoS tags and syntactic dependencies; then, to test the performance across domains, they let domain experts compose notes relating to different clinical specialties, such as cardiology and surgery. As expected, the shift of domains causes a drop in performance; overall, though, the results are relatively robust, which indicates the portability of the model.

For the *recognition of medical entities*, the commonly used processing suites for English draw heavily on existing knowledge sources. As the terminological coverage for most other languages remains poor, though, the portability of such toolkits is limited: Chiaramello et al. (2016) assess the performance of MetaMap to identify disorder mentions in clinical Italian and to link them to their corresponding UMLS concept. Due to the limited term base of the Italian version of the UMLS, the recognition rate hovers around 50%; much better results (75%) are achieved when the text is first translated into English with a generic translation tool, and then processed by MetaMap, using the full English version of the term base. Due to such limitations, most studies build custom NER systems from scratch, especially if they are geared towards specific tasks. For instance, Eriksson et al. (2013) present a knowledge-based system for the recognition of ADEs in Danish clinical notes. The recognizer relies mainly on a custom dictionary, which is based on the product descriptions supplied by pharmaceutical companies. To handle lexical variants and disjoint entities, simple rules for inflection, spelling alternation and post-coordination are implemented. In a study on clinical Swedish, Skeppstedt et al. (2014) use supervised ML to identify entities from

four semantic groups (Disorder, Finding, Pharmaceutical Drug and Body Structure) and map them to SNOMED CT concepts. They annotate a corpus of ED reports and use it to train a CRF classifier. To enrich the Swedish version of SNOMED CT, a custom term list is compiled. In addition, the recognition of term variants is improved by lemmatization and compound splitting. Pérez et al. (2017) experiment with clinical NER in a semi-supervised setting. They compare the performance of different ML approaches for the recognition of entities from various semantic classes, including drugs, diseases and body parts. Two clinical corpora, one in Swedish and one in Spanish, are partly annotated with a subset of SNOMED CT concepts. The system is evaluated with different feature combinations, including linguistic features derived from the annotated data, and conceptual features, which are obtained by unsupervised means from the un-annotated part of the corpus. While the relative contribution of the different feature types differs across languages, the results indicate that features obtained by unsupervised techniques, such as word embeddings, could partly compensate for the lack of annotated data. In a study on Belgian Dutch, Scheurwegs et al. (2017) come to similar conclusions: They compare different methods to assign ICD codes at the document level based on the mentions of diagnoses and procedures. Their dataset contains notes from multiple clinical specialties (e.g. surgery and radiology). Given the limited coverage of Dutch knowledge sources, they construct concept representations from free text. Then, they compare the impact of features obtained by supervised and unsupervised strategies for the recognition of diseases and procedures. Interestingly, the performance of the techniques differs across semantic classes: In general, distributional features are more informative for the detection of diagnoses than for procedures. Possibly, this difference can be attributed to the properties of the underlying concepts: Procedures typically involve a well-defined operation, which can be expressed by concrete terms; conversely, expressions relating to diseases tend to be vague or even speculative in nature, which increases the potential for term variation. Thus, distributional features, which are agnostic to the surface form of the target term, are particularly suited for the task.

Finally, for the interpretation of contextual properties, such as *factuality* and *temporality*, many studies rely on a knowledge-base built from scratch. For

instance, Thomas et al. (2014) develop a recognizer to assign ICD codes to a mixed sample of notes from a Danish psychiatric hospital. To identify negated instances, the system relies on a manually compiled list of cue words. This simple method produces fair results for the data under investigation. However, as its accuracy depends crucially on the proximity of the lexical trigger to the target entity, it might not be reliable for the classification of more narrative document types, such as discharge summaries, which tend to contain more complex syntactic constructions. Alternatively, the adaptation of tools developed for English can be an efficient solution, especially for languages of the same family. Afzal et al. (2014) present a Dutch extension to ConText (Harkema et al. 2009), a rule-based algorithm to determine three context features (negation, temporal order and identity of the experiencer). For development and evaluation, they use a heterogeneous data sample, including notes composed by clinicians as well as general practitioners, and clinical discharge letters. To adapt the algorithm, they translate the list of lexical negation cues and adapt the rule-base to cover negation processes in the Dutch language. In addition, the algorithm is enhanced with regular expressions to handle idiosyncratic ways of negation, which are commonly used in individual data samples. While the system struggled with the more complex task of determining the temporal order, it produced robust results for factuality and experiencer. However, when dealing with unrelated languages, the adaptation of existing tools is usually not a viable option. Based on an annotated corpus of clinical Chinese, Zhang et al. (2016) develop a system for the detection of speculation, i.e. instances of uncertainty or vagueness. As the lexical cues associated with speculation are quite ambiguous, heuristics are ill-suited. Instead, a CRF-based model is trained for the task, using embeddings at character and word level as features. For the identification of temporal details and relations, existing tools, such as HeidelTime (Strötgen and Gertz 2010), have been adapted to the clinical domain. HeidelTime is a rule-based temporal tagger, which supports multiple languages. To tune it to the clinical domain, Hamon and Grabar (2014) enrich the model with domain-specific expressions in English and French; in addition, they enhance it with functions to compute relative and approximate values. De Azevedo et al. (2018) use HeidelTime to detect and normalize temporal expressions in Brazilian Portuguese. For adaptation, they extend the knowledge-base with

standard terms as well as noisy variants identified in the training data, such as misspelled or wrongly segmented forms.

1.4 Conclusion

Clinical data science has received an upsurge of interest in the past decades; so has clinical NLP as the key technology to make this data accessible. However, while the potential benefits of data reuse, both to support healthcare operations, and to serve as input for scientific advances, are now widely accepted by the research community, few applications have found their way into clinical practice yet: In fact, most systems are "explored, published, and shelved" (Demner-Fushman and Elhadad 2016, 231).

One reason for the slow adoption might be that clinicians lack the familiarity with NLP to employ it for routine tasks. Similar to early versions of the EHR, most applications do not come in a ready-made package that can be deployed directly and customized according to daily needs; instead, their usage requires additional technical expertise, which is not compatible with the educational background and workload of many clinicians. To enable a smooth integration into daily practice, more user-friendly applications are needed. Secondly, the implementation of NLP methods needs to be geared towards pragmatic needs. Many tasks do not require a thorough linguistic analysis, but can be performed by relatively simple, easy-to-fix methods. For instance, knowledge- or rule-based algorithms have been found to perform well for many tasks, such as the recognition of particular entity types (e.g. Ford et al. (2016); Koleck et al. (2019)). Besides, using transparent methods might also lower the threshold to use them among practitioners with little experience in NLP (Wang et al. 2018). If more complex approaches are used, though, they should strike the right level of granularity: For example, the billing of clinical care is usually based on all services delivered during a patient's hospitalization, not those mentioned in an individual report (Scheurwegs et al. 2017). Therefore, with regard to financial administration, a NER system that ensures reliable recall over a set of documents might be more valuable than one that achieves high precision at word level. Thirdly, such applications must be

mature enough to produce robust results in practice. Many state-of-the-art systems are developed and tested on datasets that have been compiled for the purpose of academic research; however, their performance in a real-life setting, using heterogeneous samples of clinical data, is never evaluated and reported. Especially for under-researched languages, the currently available solutions still lack the maturity for daily use. To ensure robust performance, a solid methodological framework is required. The improvement of foundational methods, as well as the construction of high-quality terminological resources, should thus remain a priority, in particular for languages other than English.

To tackle these challenges, clinical NLP employs mostly rule-based methods or supervised ML. Faced with the particular properties of the clinical genre, though, both approaches have their drawbacks: Rule-based methods are typically easier to implement and come with a lower computational cost. Especially for tasks at the lexical level, like morphological analysis, they can be powerful tools. However, they strongly depend on the quality and scope of pre-compiled knowledge sources. Given the high degree of variation in clinical writing, the creation and maintenance of exhaustive terminologies or grammars is a demanding task. On the other hand, supervised ML relies on annotated data as input. To obtain meaningful results, the linguistic features of the training data should overlap with those of the data under investigation. However, clinical specialties and document types vary considerably with regard to their lexical, syntactic and semantic features. For optimal results, the creation of representative datasets for every clinical sublanguage would be required. Considering the costs associated with the creation of annotated corpora, this is difficult to realize, especially in minor languages, where resources are particularly sparse. Therefore, many systems opt for a compromise, combining knowledge and data in a complementary fashion to solve the task at hand.

For Belgian Dutch, there are practically no annotated datasets available for research. Structured terminologies, too, are still in the making: The first Belgian translation of SNOMED CT was released in 2018, comprising both a French and a Dutch module. In total, 26,814 concepts were translated to both

languages, amounting to 7.86% of those included in the international version. Each of these concepts is represented by one standard term. In addition, 1.350 synonyms were added to the Dutch extension, and 1.540 to the French one. However, the percentage of translated terms varies across the semantic classes: While, for instance, the category of geographical names was translated completely, the coverage of some of the medically relevant classes is only rudimentary (e.g. 2.47% for concepts belonging to the category of findings, and 1.57% for procedures; cf. Mertens (2018)). Therefore, at this point, the Belgian release is not comprehensive enough to be a functional component of the clinical workflow. Besides, as previous studies have shown that the standard terms included in translated terminologies are not representative of term usage in practice (Henriksson et al. 2014; Skeppstedt et al. 2014), it is not a sufficient knowledge source for NLP applications.

As the example of Belgian Dutch illustrates, the discrepancy between terminological resources and term usage remains a major challenge for the processing of clinical language. To bridge this gap, earlier research proposed the development of comprehensive processing systems relying on customized term bases (Spyns 2000). However, for a low-resource language like Belgian Dutch, the creation and maintenance of an exhaustive terminology seems too laborious in practice. For a more efficient approach, this thesis proposes to tackle the issue at a more abstract level: Rather than attempting to create a full inventory of all terms in usage, it will focus on a confined domain to characterize term variants by formal features, and identify the underlying variation processes. By going beyond the purely lexical level, it aims to provide a more generalizable view on term variation.

Chapter 2: Theories of Terminology and Variation

The study of terminology is a relatively young discipline of scientific research, which was only established in the 1950s. However, even in this short history, it has undergone some fundamental paradigm shifts, which crucially affect the analysis and modeling of term variation. Therefore, the first section of this chapter (Section 2.1) sketches the cornerstones of terminological theory, starting from the prescriptive theory proposed by one of the founding fathers of terminology, Eugen Wüster (Section 2.1.1), up to the descriptive approaches presented since the 1990s (Section 2.1.2). The more recent approaches offer different typologies of variation phenomena; thus, the following section (Section 2.2) gives an overview of different types of variation observed in the surface form (Section 2.2.1), and the causes that can be linked to these phenomena (Section 2.2.2). The next section assesses how term variation is represented in SNOMED CT, one of the major biomedical terminologies used for clinical coding today (Section 2.3). Finally, the findings will be summarized in the conclusion (Section 2.4).

2.1 Theories of Terminology

2.1.1 Wüster and the General Theory of Terminology

To a great part, terminological practices today are shaped by the pioneering work of Eugen Wüster. In 1931, Wüster, who had a background in engineering, published his doctoral thesis about the international standardization of the terminology of electronics. In the wake of this publication, the International Organization for Standardization (ISO) founded a technical committee that is solely concerned with the standardization of terminology and the

exchange of specialized information. For the remainder of his career, Wüster continued to develop his theories and established terminology as a separate discipline of research and academic study (Lang 1998). Among the milestones of his lexicographic work is a multilingual dictionary of machinery, which he compiled on behalf of the European Economic Committee (Wüster 1968). His theory was further developed by his academic followers and became known as the *General Theory of Terminology (GTT)*.

Wüster's work was driven by the ambition to optimize the exchange of specialized information across languages. To this end, he aimed to eliminate the sources of misinterpretation, both at the conceptual and the lexical level. His theory is based on the structuralist assumption that a concept exists as an abstract mental entity, which is referred to by an arbitrary signifier. A concept is considered a stable unit which does not change over time; the extension of a concept thus subsumes all potential instantiations. The primacy of the concepts requires that an entity be clearly defined and delineated before it can be labeled by an unequivocal term. Contrary to general language, specialized language does not allow for ambiguity or diachronic changes: Ideally, terms fulfill the criterion of mononymy and are not modified over time (Felber 1979; Wüster 1979). Terminology is thus no longer considered a branch of linguistics, but "an autonomous discipline the object of which are no longer terms considered as units of natural language, but concepts considered as clusters of internationally unified features which are expressed by means of equivalent signs of different linguistic and non-linguistic systems" (Cabré Castellví 2003, 167). Further developments attenuated some of these rigid propositions, allowing for controlled synonymy and the integration of new terms. However, the fundamentals of the theory, in particular the strict concept-orientation and the strife for standardized expressions, stayed untouched.

For decades, the GTT remained the sole theoretical framework in terminology. While the number of alternative theories is growing (cf. Section 2.1.2), the influence of the GTT on terminological policies and knowledge management remains evident in numerous specialized domains.

2.1.2 Descriptive Approaches to Terminology

Starting from the 1990s, the study of specialized languages underwent a significant shift in paradigms. Fostered by the development of automatic methods for the analysis of domain corpora, descriptive approaches to terminology began to emerge. By analyzing specialized language in practical contexts, these approaches found that the strictly onomasiological, synchronic perspective proposed by Wüster is insufficient to account for the usage and understanding of terms in practice. Instead, terms must be studied in interaction with pragmatic and cognitive factors.

Moving away from a prescriptive approach, *socioterminology* considers terms as dynamic signifiers, which are negotiated by the speech community (Gaudin 1993, 2005). Terms originate in practice and are consolidated in usage; mutual understanding is thus only enabled by conventionalization. Socioterminology proposes a dynamic perspective, where the pairing of terms and meaning is neither unequivocal nor stable. Instead, term usage reflects the background and objective of the interlocutors. Term choices might be subject to subconscious influences (e.g. to signal expert status), or to deliberate instrumentalization (e.g. to convey an ideological stance). Term variation is thus regarded as the product of changes in discourse context. Therefore, the study of terms must take the circumstances of their appearance into account, in particular by analyzing verbal interactions between domain specialists. Corpus linguistics provides the necessary toolkit for such analyses. While later interpretations of socioterminology do not preclude language planning per se, they still reject the top-down normalization envisioned by the GTT. Instead, terminologies should be compiled in a bottom-up approach, paying respect to the attitudes and practices of the users.

The *Communicative Theory of Terminology* (CTT) views terms as discourse units, which are instantiated by the interlocutors in a specialized domain. The CTT rejects the clear separability of the general and the specialized, both with regard to the underlying knowledge and its linguistic expression. Terminological units may overlap with expressions from general language; their specialized meaning is only activated by usage in domain-specific discourses.

Such discourses can be modeled by conceptual maps, where the terminological units act as representatives of nodes of knowledge. Contrary to the Wüsterian ideal, such discourses are marked by redundancies and variations, which can be attributed to differences in conception and register. Therefore, the CTT questions the principles of terminological monosemy and stability. Concepts are characterized as essentially multidimensional; term variants give access to different facets. Depending on their communicative intention, speakers may purposefully select different variants to emphasize particular nuances. Just like general words, terms have a cognitive, linguistic and sociocommunicative component; they do, however, fulfill certain conditions with regard to these components. For example, in a structured representation of domain knowledge, specialized terms can be situated at a precise spot; they convey a fixed meaning within that domain, and are used and disseminated by domain experts. Terminological analyses may choose to focus on one of these components; however, it is not possible to study them in isolation. For example, the description of linguistic forms must also be linked to the communicative function of the terms (Cabré Castellví 2003, 1999).

The most recent theoretical approaches to terminology break completely with the structuralist assumption that form and meaning can be divorced from each other. Drawing on the propositions of cognitive linguistics, they assume that language shapes knowledge; therefore, terminology should not take the concept as a starting point, but rather its lexical manifestation, and the way that language users make sense of it.

Sociocognitive terminology introduces units of understanding as a way of structuring specialized information. Temmerman (2000) argues that domain knowledge is experiential and stored in prototypical categories that depend on the perception and usage conventions among practitioners. As these prototypes aggregate exemplars that can be associated with a category, they have fuzzy boundaries. In many cases, the top-down delineation of concepts, as suggested by the GTT, enforces an artificial separation. Moreover, as illustrated by examples from the life sciences, specialized knowledge is highly dynamic and cannot be represented by a static model. With scientific innovation and scholarly renegotiation, new exemplars come into being, and

the conception of old ones may change; a category can thus widen or shift its center. On the other hand, differences in the perspective of the speaker can cause the emergence of variants with equivalent meaning. Polysemy and synonymy are thus natural corollaries of science, which can support the progress of understanding and knowledge-making; the enforced standardization of term usage is unrealistic and, with regard to scientific progress, even undesirable. This implies that all terminographic work should set out at the discourse level. Therefore, the first step towards the creation of a domain-specific term base should involve the definition of the target domain as well as the intended users and applications of the term base. Guided by these criteria, a representative corpus can be compiled, which serves as the basis for the identification of relevant terminological units. These terms are the starting point for the differentiation of units of understanding. The appropriate method for this step should depend on the properties of the unit itself: For clear-cut concepts, a traditional approach based on definition and delineation may be viable; for the majority of units, though, the evaluation of different aspects, including intra- and inter-categorical relations and diachronic developments, will be required to distill a categorical prototype. For an adequate terminographic representation, Temmerman (2000) suggests the use of customized templates. Such templates should specify the type of category (e.g. an entity, process or umbrella term) and its relation to other units. Moreover, it should include linguistic information (e.g. morpho-syntactical properties, synonyms and usage contexts), as well as bibliographic references and identifiers of the terminological record.

Termontography is a more practically oriented approach, which combines the socio-cognitive angle with knowledge engineering. To support the common understanding of specialized topics, terms acquired from corpora are integrated into an ontological model. To guide the terminographic procedure, termontography proposes the a-priori development of a categorization framework. This framework specifies the formal and semantic inclusion criteria of relevant terms, as well as relevant information for ontological modeling, such as meta-categories, hierarchical and associative relations. For continuous improvement, the ontology might be enriched with more fine-

grained properties and relations, or aligned with existing knowledge sources (Faber 2009, 2012; Kerremans, De Baer, and Temmerman 2010).

Frame-based terminology (FBT) is based on the observation that specialized texts tend to adhere to conventional structures to fulfill their communicative purpose. Such structures can be exploited to identify the central events of a domain and to craft structured templates for their representation. These templates can be further refined to represent more specific units and fine-grained semantic relations. Crucially, though, they cannot be transferred across knowledge areas, but must be designed with respect to domain-specific entities and processes. FBT is similar to other descriptive approaches in that it emphasizes the importance of usage-based methods to acquire and structure specialized terms. Just like socio-cognitive terminology, it also strives for a representation of knowledge that reflects human cognition. Contrary to the theories presented earlier, though, FBT combines top-down and bottom-up methods: To structure a specialized knowledge field, FBT relies on both existing resources and input from domain experts, as well as the study of domain corpora. For an adequate representation of individual terms, FBT aims for a thorough representation of their multidimensional nature. Multi-modal terminologies can support this goal, for instance by providing graphical illustrations (Faber et al. 2007; Prieto Velasco and Tercedor Sánchez 2014). Moreover, FBT proposes to enhance terminological units with information about their syntagmatic relations. Drawing on the propositions of Frame Semantics (Fillmore 1992), FBT assumes that, in order to communicate about a specialized field of knowledge, one needs to be aware of the semantic relations between concepts, but also the conventional term combinations and appropriate situations of use (Faber 2009, 2012).

2.2 Typologies of Term Variation

Daille et al. (1996) define a term variant as “an utterance which is semantically and conceptually related to an original term” (ibid., 201). In other words, a variant is a form encountered in discourse, which diverges from the reference standard, but can still be linked back to it through semantic or

linguistic processes. Depending on the theoretical framework, variation types have been classified either by *cause*, i.e. the factors triggering the variation process, or *effect*, i.e. the properties or alternations observed in the form. In line with the logic of descriptive terminology, which takes the forms observed in text as a starting point, the following sections first characterize the different manifestations of variation in the surface form (Section 2.2.1), and then give an overview of the factors that pattern with particular types of variation (Section 2.2.2).

2.2.1 Formal Types of Variation

For the formal characterization of term variants, a number of typologies have been proposed. The dimension and granularity of these classification schemes depends largely on the intended application: For instance, a terminology used for machine translation might employ other criteria than a terminology used for bibliographic indexing (Daille 2007). For the purpose of this thesis, the typology proposed by Daille (2018) will be adopted. However, with an eye on the practical application, i.e. the general description of variation types in the clinical domain, some fine-grained distinctions will be omitted.

Thus, the following sections outline variation processes at three levels: *Conceptual variation* involves a divergence between the conceptual properties of the variant and those of the referent. *Denominative variation* relates to processes that preserve the semantics of the referent, but result in a lexicalized form whose differences with regard to the base form cannot be attributed to regular alternations. Both conceptual and denominative variants can cause *linguistic variation*, i.e. the creation of a different surface form, in order to comply with structural requirements of the language system, such as grammatical rules or spelling conventions.

2.2.1.1 Conceptual Variation

Conceptual variation creates a conceptual discrepancy while preserving a close semantic relation; hence, the resulting variant can be characterized by its relation with the referent.

Among the elementary types of inter-conceptual relations are *synonymy* (complete or nearly complete equivalence of meaning), *antonymy* (opposite meanings), *hypernymy* (one meaning subsumes the other) and *meronymy* (one meaning is part of the other; cf. Cruse (1986)). While the last three phenomena relate to variants in the sense of the definition, synonyms are often borderline cases, as their semantic configuration overlaps largely with that of the referent. At closer analysis, though, the meaning of two variants is rarely completely identical. Ultimately, the border between synonyms and related concepts cannot be defined universally, but depends on domain-specific aspects, and, in applied contexts, on the intended purpose of the task at hand. However, just like antonymy, cases of near-synonymy are rarely stated explicitly in terminological resources. Hypernymy, on the other hand, plays a prominent part in the structuring of specialized knowledge fields: Hierarchical relations are not only crucial for the structuring of a knowledge field, but also shape the implementation of ontological resources, including the inheritance of conceptual properties (Faber and L'Homme 2014). Meronymic relations, on the other hand, are more challenging to model: As Sager (1990) points out, the set of relevant relations (e.g. *cause – effect, activity – place*) depends largely on the domain, and the semantic properties of the entities involved. In many specialized languages, meronymic relations are expressed by conventionalized phrases, which can be exploited to identify such relations, and the types of concepts that typically take part in them (Meyer 2001).

According to Daille (2018), the primary processes involved in conceptual variation are *expansion* and *reduction*.

Term expansion involves an elaboration of the surface form, which results in a refinement of the conceptual structure. Expansion is often achieved by *derivation* (e.g. *embolus – embolism*). Depending on the flexibility of the morphological system, term expansions can also be created by predication, i.e. the insertion of the base form into the argument structure of another term (e.g. *sentinel node – sentinel node biopsy*), or by modification, i.e. the addition of another lexical element at the beginning, middle or end of the term (e.g. FR *diabète insulino-dépendant* 'insulin-dependent diabetes' – *diabète insulino-non-dépendant* 'non-insulin-dependent diabetes').

In cases of term reduction, on the other hand, a lexical element is omitted, resulting in a more generic conceptualization. *Anaphoric reductions* are the result of a discursive process, whereby a constituent of a previously mentioned term is left out to avoid repetition (e.g. *enzymatic process – process*). By contrast, in *lexical reductions,* the omission takes places without prior mention of the full form; the intended sense can only be inferred by domain knowledge.[1]

Apart from the mechanisms described by Daille (2018), conceptual variation can be caused by *permutation.* In many complex terms, such as nominal compounds (e.g. NL *diabetesretinopathie* 'diabetes retinopathy' (own example[2])) or terms consisting of multiple neoclassical confixes (e.g. *otorhinolaryngology*), the relation between the constituents is underspecified. However, as Bowker and Hawkins (2007) argue, the sequence of elements can convey different semantic nuances, such as the directionality or etiology of a medical finding (e.g. *cardiovascular – vasculocardiac*). Similarly, the order of names in complex eponyms can reflect the chronological order of scientific advances (e.g. *Klippel-Trénaunay syndrome*, which, following the scientific contribution of Weber, is now also known as *Klippel-Trénaunay-Weber syndrome*).

2.2.1.2 Denominative Variation

While denominative variation preserves the semantics of the referent, the resulting form may carry out a different communicative function, e.g. to enable the communication with non-specialists. Denominative variation can create both fully lexicalized forms, such as conventionalized abbreviations, and syntactically flexible expressions, such as paraphrases. Daille (2018) identifies three processes that create denominative variation, namely *substitution*, *simplification* and *exemplification*.

[1] The question of whether the non-anaphoric omission of lexical elements is an instance of conceptual or denominative variation has been subject to some debate. Daille (2018) argues that, as the dropped elements are typically not essential to characterize the concept, lexical reductions produce variants that are semantically equivalent. Likewise, in this thesis, lexical reductions will be situated at the denominative level, together with other reduction processes.

[2] All examples in Dutch (prefixed by NL) are taken from the dataset presented in Chapter 5 of this thesis.

Substitution relates to the replacement of one or multiple components by equivalent lexemes. In the medical language, substitutions frequently involve alternations between native lexemes and neoclassical elements, resulting in a change in register (e.g. NL *cor – hart* 'heart'). Likewise, words from modern foreign languages, in particular English, can be inserted or dropped from native terms. Another domain-specific process is the replacement of eponyms, i.e. terms based on personal names, with terms from the scientific nomenclature (e.g. *Eustachian tube – auditory tube* (Wermuth and Verplaetse 2019)).

Simplification is a highly productive process in the medical domain, especially in clinical communication. It is typically achieved by the compression of the term to a reduced form, as it is the case with acronyms (e.g. *HIV* 'Human Immunodeficiency Virus' (Wermuth and Verplaetse 2019)). Alternatively, lexical or functional elements can be dropped altogether (e.g. NL *CT scan – CT*; *rx van de thorax – rx thorax* 'x-ray of the thorax').

By contrast, exemplification makes the semantic structure more explicit. This can be achieved through the insertion of additional lexemes that support a term's conceptualization (e.g. *dark urine – dark-colored urine*), or through the decompression of term structure (e.g. *breast cancer treatment – treatment for breast cancer* (Daille 2018)).

2.2.1.3 Linguistic Variation

Linguistic variation refers to changes in the surface form, which are required to instantiate conceptual or denominative variants in a language system; this can involve processes at the *syntactic* and *orthographic* level.

Syntactic variation is highly systematic, as it is mostly an effect of grammatical constraints. With regard to *inflection*, the number of potential variations is determined by the richness of the morpho-syntactical system. Besides these regular alternations, though, there are a number of phenomena which are less predictable. In particular, the inflection of foreign words or neonyms, whose morphological properties do not fit into the native system, can result in additional variation. In the medical domain, this is typically an effect of the

use of terms based on neoclassical languages, whose grammar differs from that of the native language. Especially in Germanic languages, such terms tend to be inflected according to the rules of the source language (e.g. DE *Arteria coronaria* 'coronary artery' – *Arteriae coronariae* 'coronary arteries' (Wermuth and Verplaetse 2019)). *Derivation* creates new forms based on the lexical core of an existing term (e.g. FR *insulino-résistance* 'insulin-resistance' – *insulino-résistant* 'insulin-resistant' (Daille 2018)). In the medical domain, which is dominated by nominal constructions, the most frequent type is the derivation of adjectives, which are then used as modifiers in complex noun phrases.

Due to the high proportion of foreign elements, orthography is an important factor of variation in medical language. For instance, for many terms based on Greek roots, different *spelling variants* are acceptable in the native language. Spelling variation can also spill over into the orthography of reduced forms (e.g. DE *Electrocardiograph – Elektrokardiograph, ECG – EKG*). However, the majority of these variations follow regular patterns. By contrast, *accidental misspellings,* which are particularly frequent in clinical communication, are thoroughly unpredictable (e.g. *isnuline – insuline* 'insulin'). Finally, the use of *spaces* or *hyphens* in complex terms can produce different surface forms. In Germanic languages such as German and Dutch, compound forms are written as one word (e.g. DE *Herzschrittmacher* 'pacemaker') or joined by hyphens (e.g. DE *Billroth-I-Operation* 'Billroth's operation I' (Wermuth and Verplaetse 2019)). However, the native conventions are increasingly replaced by the pattern found in the English language, whereby the constituents are separated by spaces (e.g. NL *insulineinjectie – insuline injectie* 'insulin injection').

2.2.2 Causes of Variation

2.2.2.1 Conceptual Variation

Conceptual variation can be caused by changes in the physical reality, or the state of knowledge. With the discovery of unknown entities and phenomena, and the development of new techniques, lexical gaps appear in the vocabu-

lary of a specialized language. Neonymy serves to provide an adequate linguistic representation for such instances, either by refining existing expressions or by coining entirely new terms (Roldán Vendrell and Fernández-Domínguez 2012).

Besides, conceptual variation can be caused by cognitive factors. On the one hand, scientific advances and practices can lead to changes in the categorization of known entities; accordingly, the core sense of an existing term can shift, expand or narrow (Cabré Castellví, Bagot, and Vargas-Sierra 2012). On the other hand, differences in the perception and interaction with an entity, or the role a speaker takes in a process, can influence the conceptualization (Tercedor Sánchez 2011; Prieto Velasco and Tercedor Sánchez 2014). In clinical communication, which involves both laypeople and domain specialists, the variation potential is increased by differences in the knowledge level. Especially with concepts that have no outward manifestation, such as symptoms of pain, variation can occur at two stages: firstly, during the subjective evaluation by the experiencer (i.e. the patient); and secondly, during the association with a medical concept by the observer (i.e. the doctor). By contrast, concepts that can be assessed by their physical appearance, such as anatomical entities, tend to be more stable (Wermuth and Verplaetse 2019).

Finally, conceptual variation can be caused by geographical, dialectal and institutional influences. In many scientific disciplines, the preferences conveyed by academic education (i.e. a particular "school of thought") can have a lasting effect on the understanding and expression of specialized knowledge (Bowker and Hawkins 2007).

2.2.2.2 Denominative Variation

In many types of professional communication, denominative variation is a product of pragmatic circumstances. In clinical writing in particular, the influence of the fast-paced working environment is evident. As many clinicians work under extreme time pressure, the use of concise expressions is crucial, even at the expense of transparency. The priority of speech efficiency is most evident in the high proportion of reduced forms and short jargon expressions. Moreover, differences in the knowledge level of the interlocu-

tors can cause variation. For mutual understanding, clinicians may use lay variants when consulting with their patient; the decrease in specialization tends to increase the potential for variation (Cabré Castellví 2003). The opposite phenomenon, which is known as "perverted adequacy" (Freixa 2006, 57) can also occur: In sensitive situations, doctors may purposefully employ a language that makes it difficult for their patients to follow. Likewise, in the internal exchange between peers, slang and jargon forms might be more frequent.

Denominative variation can also be due to interlinguistic effects. In particular, English, which is the dominant language of scientific communication today, exerts an increasing influence on specialized communication in other languages. This effect is most evident in the use of direct loans. However, there is also evidence for indirect influences on term formation and usage. For instance, it has been found that, in Spanish, the argument and event structure of neonyms in the domain of Alzheimer shows strong parallels to the English equivalents (Ibáñez and Palacios 2014).

Finally, denominative variation can be caused by discursive factors. In biomedical writing, new terms may be introduced to avoid repetitions, or to emphasize the originality of an approach (Cabré Castellví, Bagot, and Vargas-Sierra 2012; Daille 2007). While stylistic considerations are of lesser concern in the clinical genre, such variants can, eventually, infiltrate the vocabulary used in clinical practice.

2.2.2.3 Linguistic Variation

Linguistic variants serve to instantiate conceptual or denominative variants in a way that complies with the language system. Thus, this type of variation is not an effect of the domain, but rather of grammatical constraints and spelling conventions. However, whether such rules are followed or not depends mainly on pragmatic circumstances: While biomedical pieces of writing, such as scholarly articles, are composed in the well-formed style required for publication, clinical notes tend to be composed in a telegraphic style which defies grammatical conventions. In fact, it is the absence of regular morpho-syntactical alternations that is characteristic for the clinical genre. At the

same time, the hectic environment of the clinical setting may increase the rate of accidental misspellings, and increases the tendency to omit punctuation.

On the other hand, systematic orthographic alternations, as they occur with neoclassical terms, can depend on geographic or institutional conventions, as well as individual preferences.

2.3 The Representation of Term Variation in Medical Knowledge Sources

As illustrated by the previous sections, the past two decades have seen an increased interest in the documentation and classification of term variants encountered in usage. Gradually, the cognitive shift in terminological theory has raised awareness of the shortcomings of the design principles proposed by the GTT with regard to the construction of terminological resources. As a consequence, a number of projects attempted to construct more adequate representations of biomedical language, both by enabling flexible ontological structures, and by including rich context models. For instance, Cabré Castellví et al. (2004) present GenomaKB, a knowledge-rich terminology for the genomic domain. This knowledge base combines different modules to represent ontological structures, including fine-grained semantic relations (e.g. cell replications), terms and variants, and corpus-based information on term frequencies and usage contexts. Depending on the application, the user can prioritize the most relevant type of information. Bousquet and Zimina-Poirot (2010) present a trilingual term base of ADEs. The PERTOMed terminology is based on aligned corpora in English, French and Russian; the term entries are enhanced with abbreviations and multi-word expressions (MWEs), i.e. more complex terminological units containing the base terms. Similarly, Sambre and Wermuth (2010) employ a bilingual corpus of bio-medical text from the domains of surgery and cardiology to develop a framework for the fine-grained representation of instrumentality. In the medical domain, instrumentality can be an important cue to infer causal chains, such as *diagnosis – treatment – outcome*. However, the authors observe that most terminologies only reflect a limited number of non-

hierarchical associations and only express the links between individual concepts. For a more adequate representation, they develop a typology of instrumentality, including implicit relations (e.g. between constituents of compounds), as well as relations encoded by non-nominal forms. Marshman (2014) presents a bilingual terminology of the domain of breast cancer to support the translation of specialized texts. As she points out, knowledge-rich contexts are a very efficient way to exemplify semantic and lexical relations, which are underrepresented in structured terminologies. By exploiting lexico-semantic context patterns, she enriches the entries of her term base with semantic relations, as well as the lexical markers that are habitually used to express them.

However, even though such noteworthy initiatives exist, they remain relatively fragmented; typically, they are realized within individual projects, and only cover a limited medical field. Conversely, the more comprehensive biomedical terminologies like SNOMED CT (International Health Terminology Standards Development Organisation (IHTSDO) 2019) strongly adhere to the principles of traditional terminology: SNOMED CT is strictly concept-oriented and has a hierarchical structure. It does, however, allow for the modeling of polyhierarchial relations; one concept can thus be associated with multiple hypernyms, reflecting different classification schemes or methodological approaches. Each concept is qualified by a set of defining characteristics, and, optionally, attributes and qualifying characteristics; the range of applicable properties depends on the semantic category. For instance, a concept relating to a procedure can be refined by an attribute specifying the method, while an anatomical concept cannot. With regard to terminological units, SNOMED CT follows a rather prescriptive scheme: Each concept is associated with several descriptions, including at least a fully specified name (FSN), a preferred term (PT), and, optionally, one or more synonyms. Each term must fulfill the quality criteria specified in the editorial guidelines: In particular, the FSN is supposed to provide a unique lexical referent, which is stable across contexts. Conversely, the PT is a term that is commonly used in clinical practice or research. As term preferences may vary across the dialects, the PTs of one concept may differ across the reference sets of one language (e.g. the Australian and British English version of

SNOMED CT). The PT of one concept may also be a synonym of another concept, if its meaning can change in context. Synonyms are additional acceptable variants; just like PTs, they can be polysemous terms, i.e. they can be linked to multiple concepts.

Overall, SNOMED CT does implement some of the propositions of descriptive theories of terminology, both with regard to its ontological structure and the representation of terminological units. The possibility to specify polyhierarchical associations allows for the inclusion of multiple ways of conceptualization; this is in line with socio-cognitive theories, which claim that the taxonomical positioning of concepts must be flexible. Moreover, the enrichment with concept-specific attributes provides guidance about the constraints governing a term's combinatorial value at the semantic level. Still, at the terminological level, SNOMED CT follows a prescriptive approach: As normalization has priority, term variants are ranked by formal standards. While the use of reduced forms is discouraged, non-canonical forms, such as informal jargon, are not included at all. Another drawback is the complete lack of context information: While the specification of term types gives an indication of term preferences in practice, the concept entries do not provide concrete examples of usage contexts.

2.4 Conclusion

Even though the study of terminology is a comparatively young discipline of research, it has already undergone some fundamental paradigm shifts. Early theories, in particular the GTT, aimed primarily at the standardization of specialized language. However, as illustrated by the analysis of domain corpora, the complete normalization of specialized knowledge and its linguistic expression is an unrealistic goal. Rigid ontological structures cannot reflect the conceptual fluctuations in specialized domains; moreover, the prescriptive approach runs counter to intuitive term usage. The top-down implementation of standardized forms limits the expressiveness of natural language and might, in the end, impede communication rather than facilitate it.

Notably, it has been pointed out in the early days of the EHR already that the reliance on knowledge sources created in a top-down fashion may impede the electronic management and processing of health data (Cimino 1998; Rector 1999). Still, SNOMED CT, a comprehensive biomedical terminology, which is routinely used in clinical practice, only provides a limited reflection of term variation and does not give details on term usage. Bodenreider, Smith, and Burgun (2004) see the reason for this discrepancy in the historical development of biomedical terminologies: While they were originally designed to serve as reference standards for physicians, or to support bibliographic indexing, the scope of applications making use of them has widened considerably with the rise of NLP. At the same time, no systematic assessment of the structural changes required for these applications has taken place. Consequently, clinical NLP relies heavily on resources that are ill-prepared to meet its needs. To make the existing terminologies better suited to NLP applications, the inclusion of terms encountered in usage would be required. Moreover, the research community would need to agree on a set of descriptors for the classification of term types, and to elaborate a framework for the representation of context parameters.

Chapter 3: The Sublanguage Properties of Clinical Language

The sublanguage theory proposed by Zellig Harris has inspired the analysis and processing of a wide range of specialized languages, including the medical one. Therefore, this chapter takes a closer look at the sublanguage behavior of clinical writing. The first section summarizes the theoretical foundations of the model and sketches the sublanguages properties of clinical language (Section 3.1); the second section gives an overview of earlier descriptions of sublanguages in the clinical domain, as well as NLP applications inspired by the theory (Section 3.2). To conclude, the final section assesses how sublanguage theory can support clinical NLP in dealing with term variation (Section 3.3).

3.1 Theoretical Foundations

3.1.1 The Sublanguage Theory of Zellig Harris

The language model proposed by Harris is strongly rooted in mathematical principles and Information Theory: "All occurrences of language are word-sequences which satisfy certain combinatory constraints; furthermore, for reasons related to mathematical Information Theory, these constraints express and transmit information" (Harris 1991, 5). The way language systems operate and encode meaning is thus governed by a set of fundamental constraints: Firstly, the functioning and meaning of words is determined by their *dependency relations*. According to the strength of dependence on other words, they can be assigned to different classes: Zero-level words can be used in isolation, while higher-level words require the presence of other words to result in a legitimate utterance. For example, in the sentence *John arrived*, only the word *John* conveys meaning on its own; *arrived*, on the

other hand, requires the presence of a zero-level word like *John*. Secondly, words differ with regard to the probabilities of co-occurring with each other. For example, *rent* is more likely to co-occur with *room* than with *city*. Words can thus be characterized by *inequalities of likelihood*. Thirdly, only those words that provide a gain in information are essential to communication; uninformative words can be omitted by means of *paraphrastic reductions*. For example, *John took math before John took physics* can be reduced to *John took math before physics* without a change in meaning (Harris 2002, 217–18).

Compared to general language, scientific sublanguages are governed by a more restrictive set of semantic constraints. In general language, the acceptability of an utterance depends primarily on structural criteria (i.e. the selection of words from particular classes according to their dependencies). However, there are no restrictions on the words that might instantiate these classes. Therefore, non-sensical utterances like *I rented a city* might occur, as long as they are grammatically acceptable. By contrast, in sublanguages, the co-occurrence constraints are semantically loaded, reflecting the knowledge patterns and conceptual relations of the domain. First-level words do not only require the presence of zero-level words, but that of zero-level words from particular semantic classes. For example, in biochemistry, the phrase *is injected into*, i.e. a first-level construction, will combine with a zero-level word from the subset of *antigens,* rather than with one from the subset of *lymph nodes* (Harris 2002). On the other hand, since sublanguages deal with a confined subject matter, they allow for more extensive paraphrastic reductions. Words that would be required to form a valid utterance in general language can be omitted, since they are implicit in the restricted context of the domain. For example, in a report on an x-ray of the chest, the statement *infiltrate noted* would be a viable short form for *infiltrate in lung was noted by radiologist,* since both the observer (*radiologist*) and the anatomical location (*lung*) can be derived from context (Friedman, Kra, and Rzhetsky 2002, 225). Hence, while general language can be described by the inequalities of likelihood between word classes, sublanguages can be described by a fixed set of patterns involving subclasses of words (e.g. terms belonging to certain semantic types). These constraints can be expressed in formal nota-

tion, which can be summarized in *sublanguage grammars* (Harris 1982; Harris and Mattick 1988; Harris 1991; Harris 2002). However, sublanguages are not designed by a scientific authority. They are closed subsystems of general language, which arise spontaneously to enable the communication within a specialized community. Irrespectively of the country of practice, they are used under similar circumstances and serve similar purposes; therefore, the sublanguage features within a domain may show cross-linguistic regularities, regardless of whether the native languages are related (Kittredge 2003).

3.1.2 The Sublanguage Properties of Clinical Text

Sublanguage theory has served as a theoretical framework for the analysis of language use in a number of specialized domains, including law (Charrow, Crandall, and Charrow 1982), aviation (Kittredge 2003) and business communication (della Volpe, Elia, and Esposito 2018), and has also found wide application in the life sciences. In the medical domain, a central distinction is made between the sublanguage used in the *biomedical* literature and that used for the documentation of medical events in the *clinical* environment. While both can show substantial overlap with regard to their topics, they differ fundamentally in stylistic properties, depending on their communicative function and circumstances of production.

Biomedical language is composed by medical experts to communicate about scientific findings. It is thus written in the well-formed style required for academic publications and usually focused on a narrow, highly specialized subject matter. Based on the analysis of literature on a particular subfield of medical research, more detailed descriptions have been proposed (cf. Harris (2002) for the domain of cellular immunology; Sager, Friedman, and Lyman (1987) for pharmacology). The sublanguage properties of biomedical language are most evident with regard to the co-occurrences of semantic types, as the set of relevant entities and their combinations is extremely constrained. On the other hand, at the formal level, biomedical language mostly complies with the rules of general language. Like other genres of scientific writing,

biomedical language is dense with information, and tends to employ complex syntactic constructions. For example, the biomolecular sublanguage tends to use deeply nested constructions to express the interactions between different substances (e.g. *Inhibition of 4 e-bp1phosphorylation enhanced 4 e-bp1 binding to eif-4e* (Friedman, Kra, and Rzhetsky 2002, 231)). However, while such sentences may be difficult to process for a non-specialist, they do not violate the grammar of general language.

By contrast, the clinical sublanguage originates in the practical setting and deals with individual cases. While it is also a highly specialized form of peer-to-peer exchange, it is not intended for public communication. Depending on the document type, the formal constraints governing word combinations in general language can be neglected. Sentences that would be deemed ungrammatical in general, as well as in biomedical language, can be acceptable. In the clinical domain, sublanguage properties manifest themselves both with regard to the selection of words, their habitual combinations, and the dependency constraints they must satisfy. Compared to general language, the clinical sublanguage can thus be characterized by its *finiteness* (i.e. the use of a limited set of words, semantic combinations and syntactic constructions), *skewed distributions* (i.e. differences in the frequencies of words and word types compared to general language), as well as its *deviancy* (i.e. the prevalence of words and syntactic structures that only occur in this domain; cf. Kittredge (2003)).

3.2 Descriptions and Applications Based on Sublanguage Analysis

3.2.1 Descriptions of Clinical Sublanguages

A wide range of approaches has been used to characterize sublanguages by their lexical, syntactic and semantic features. Early studies mostly rely on the manual or semi-automatic acquisition of distinctive syntactic and semantic patterns. For instance, Friedman (1986) and Sager et al. (1994) start from the

observation that clinical language employs a limited set of sentence types to encode medical information (e.g. *test and result*, *treatment by medication*). Based on a sentence-by-sentence analysis of clinical documents, they identify the elementary knowledge patterns and map the dominant concepts and relations to information templates. Dunham (1986) focuses on syntactic deviancies in medical diagnostic statements. He shows that, while clinical language is dominated by complex noun phrases, the internal relations between the constituents of these phrases are often left implicit. For instance, in nested prepositional phrases (e.g. *disease with symptom at body part*), the connecting prepositions can be left out or replaced by orthographic symbols. As such reduced statements can only be resolved by domain knowledge, they are a typical sublanguage feature of clinical text. Friedman, Kra, and Rzhetsky (2002) compare the sublanguage features of clinical language with those of biomedical literature in the biomolecular domain. For each sublanguage, they provide an inventory of the dominant semantic types, as well as their typical combinations (e.g. *body part* and *location*). They conclude that, while there is some overlap in the semantic combinations, there are considerable differences with regard to the amount of details provided for individual classes.

The availability of large EHR collections in digital form enables the automatic induction of syntactic and semantic patterns from clinical corpora. For instance, Kate (2012) presents an unsupervised method to infer a sublanguage grammar from a sample of discharge summaries. Based on PoS tags and pre-assigned UMLS concepts, the system uses a cost-reduction technique to derive the minimal set of rules required to model the sentence types. Peterson and Liu (2018) also rely on data-driven methods to infer semantic patterns. They analyze a large corpus of documents containing clinical problem lists, which serve to summarize all EHRs linked to an individual case. To structure the characteristic information patterns, they first parse the text into syntactic triplets (*subject*, *predicate, object*), and then map the constituents to ontological concepts.

The large-scale analysis of EHR corpora also facilitates the characterization of clinical sublanguages by their lexical and conceptual structure. Temnikova

et al. (2013) use closure properties to validate the sublanguage status of a particular type of EHRs in Bulgarian. Their approach exploits the phenomenon that, compared to general language, sublanguages employ a limited vocabulary. In an iterative process, the rate of new words encountered in clinical text should reach a saturation point earlier than in a general language sample. The analysis confirms this trend at the level of words, as well as PoS types and sequences (cf. Temnikova et al. (2014) and Cohen, Baumgartner, and Temnikova (2016) for follow-up work). Feldman, Hazekamp, and Chawla (2016) use another information-theoretic metric, namely perplexity, to detect sublanguage properties in the vocabulary structure. As perplexity reflects the predictability of a word or word sequence in a sample, a lower value can indicate a trend towards finiteness, which is a marker of constrained domains. The study reveals differences in the vocabulary richness of EHRs from different clinical specialties, such as nursing notes and radiology reports.

A number of studies employs clustering to compare and contrast sublanguages used in different EHR types and clinical specialties. Zeng et al. (2011) study sublanguage properties in a very large corpus, comprising more than 100 EHR types. They cluster documents based on surface features, such as length and section headers, as well as vectors representing words and concepts. They obtain cohesive clusters relating to common themes, such as mental health, which provides evidence for the existence of robust sublanguage features within clinical specialties. Patterson and Hurdle (2011) cluster a clinical corpus comprising 17 different EHR types by their conceptual structure. They find crucial differences with regard to the thematic scope of the documents, which affect sublanguage properties: Types covering a broad semantic spectrum (e.g. admission history and discharge summary) cluster together, whereas documents relating to a confined specialty (e.g. radiology) form disjoint sets. Building up on this work, Doing-Harris et al. (2013) investigate whether these differences prevail in a cross-institutional comparison. They find that, even though the names and internal organization of the documents vary across hospitals, they employ a similar range of sublanguages.

3.2.2 Clinical NLP Applications Based on Sublanguage Analysis

The insights gained from sublanguage analysis have been leveraged by NLP systems addressing a range of different tasks, including vocabulary and knowledge acquisition, word sense disambiguation (WSD) and parsing.

Johnson and Gottfried (1989) point out that, in order to efficiently support experts in their workflow, knowledge sources should be constructed in a bottom-up fashion. To reflect patterns of usage, domain terminologies should be enriched with context information. They conduct a co-occurrence analysis to group the words used in a clinical sublanguage according to their combinatorial value and derive semantic patterns. The dominant patterns are summarized in formulas reflecting the dominant information structures in the field (e.g. *antibody – operator – tissue*). Zhao et al. (2018) develop a system for the data-driven acquisition of a knowledge model in medical imaging. Using existing NLP tools for the medical domain, they analyze a corpus of radiology reports at the syntactic and semantic level. The semantic types of the recognized entities are fed into a network analysis, where the dominant semantic patterns are identified by the weight of the edges.

Patterson, Igo, and Hurdle (2010) present an automatic method for the acquisition of disambiguation rules from clinical corpora. They analyze a heterogeneous sample of EHRs, which vary with regard to the clinical specialty (e.g. cardiology, dermatology), as well as the professional role of the authors (nurses and physicians). Based on a co-occurrence analysis, they show that, compared to biomedical literature, individual EHR types are dominated by narrow set of semantic patterns. These patterns can support NER applications in resolving polysemous terms.

Sublanguage grammars are at the core of numerous processing systems for the syntactic and semantic parsing of clinical text. Friedman et al. (1994) present a processing system for radiology reports, MedLEE (Medical Language Extraction and Encoding System). Drawing on earlier sublanguage analyses conducted as part of the Linguistic String Project (LSP; cf. Sager,

Friedman, and Lyman (1987)), they design a semantic grammar to map domain-specific knowledge patterns to standardized templates. These templates, so-called information formats, comprise an extensive set of modifiers for the context-rich representation of entities, concerning medical details (e.g. body location), as well as the information status of a finding (e.g. uncertainty). In subsequent work, MedLEE was further improved for the processing of other document types, such as discharge summaries (Friedman et al. 1996, 2004). The system is still widely used by the clinical NLP community today (Savova et al. 2010b).

Besides, a number of studies relies on grammar-based processing systems for the parsing of particular languages or medical specialties. Spyns (2000) presents an NLP system for clinical Dutch, which draws on a sublanguage grammar for the processing at the morphological, syntactic and semantic level. He observes that, due to the semantic constraints of the domain and similar conditions of practice, clinical sublanguages show cross-lingual universalities, for instance with regard to the formation and usage of neoclassical neonyms. Based on this finding, he proposes a modular architecture, where language- and sublanguage-specific functions are kept apart. Laippala et al. (2009) develop a parser for ICU reports in Finnish. Using the output of an existing morphological analyzer, they design a grammar based on typed feature structures. Xu et al. (2011b) present a parser for medication statements. They draw on a pre-annotated corpus to extract semantic patterns and compile a context-free grammar. To resolve syntactic ambiguities, they enhance the system with probabilities derived from a manually annotated treebank.

3.3 Conclusion

Sublanguage theory is a powerful framework for the analysis and processing of specialized languages. Its particular benefit lies in the capacity to model structural deviancies from general language, and, at the same time reflect the fact that such deviancies are not arbitrary. In fact, such phenomena are the

result of semantic constraints and circumstances of practice; consequently, the features of one sublanguage can generalize across language systems.

Insights based on sublanguage theory have thus made an essential contribution to the analysis and processing of clinical writing. The past decades have seen a trend towards more fine-grained distinctions, and the customization of NLP systems to improve their performance in individual clinical specialties. However, given the number of clinical subdomains, the fine-tuning of NLP systems to every specialty is not feasible, especially in under-resourced languages. Therefore, to strike a balance between performance and computational efficiency, sublanguages should be delineated in a more efficient way. As illustrated by recent work, clustering can serve to induce a typology of sublanguages among heterogeneous EHR types. Such a typology can serve to identify families of sublanguages that show comparable phenomena and can be processed by similar computational strategies. In addition, the transferability of clinical NLP systems can be increased by following modular design principles. In particular, the separation of language-specific functions from sublanguage-specific components would enable the development of models that are sharable across languages.

With regard to term variation, this implies a clear division between the modules handling linguistic variation (which is mostly an effect of rules governing the linguistic system), and those dealing with conceptual and denominative variation (which can be related to the domain and practice of usage). While the former can be solved based on grammatical rules in the local language, the latter must be passed on to a sublanguage-specific module, which is sensitive to variation patterns caused by cognitive and pragmatic factors.

Part II

Chapter 4: Motivation and Aims of the Case Study

The first part of this thesis illustrated the impact of term variation on the automatic processing of clinical writing, and presented theoretical proposals for the analysis of variation phenomena in specialized languages. In the second part, these theories will serve as the backdrop for a practical investigation of term variation in clinical sublanguages. This chapter thus serves as a link between the two parts: The first section elaborates on the theoretical motivation of the case study presented the following chapters (Section 4.1). The next section formulates the main hypothesis, which is at the core of this case study (Section 4.2). To give an overview of the case study itself, the final section lays out the analytical and experimental steps and explains how they tie into the general objective (Section 4.3).

4.1 Theoretical Motivation

The previous two chapters laid the theoretical groundwork for the analysis of clinical terminology in usage. As described in Chapter 2, specialized terminology is far from stable, but shows a wealth of variation processes. According to socio-cognitive theories of terminology, these processes are situated at the conceptual, denominative and linguistic level, and can be conditioned by cognitive and pragmatic factors, as well as rules of the language system. Sublanguage theory, which was introduced in Chapter 3, envisions specialized languages as linguistic systems, which are governed by semantic constraints and conditions of practice; together, these factors determine the prevalent syntactic dependencies and dominant patterns of co-occurrence. The choice of a term in a given context is thus the result of these constraints.

While both socio-cognitive terminology and sublanguage theory deal with related phenomena, they approach them from different angles, and, conse-

quently, employ different methodologies: Terminological studies typically focus on the qualitative analysis of individual forms to classify different types of variants by cause and effect (e.g. Bowker and Hawkins (2007); Daille (2018)). They describe changes in the surface form, relate them to linguistic processes and, ultimately, contextual factors that triggered these processes. However, they typically pay little attention to the significance of these processes within the language system; instead, they describe the properties of individual terminological units as a product of context (e.g. Faber and León-Araúz (2016)). On the other hand, sublanguage analysis relies primarily on quantitative techniques to characterize linguistic subsystems in their entirety. Typically, individual sublanguages are characterized by structural features, such as the proportion of word classes in the vocabulary, or formal properties, such as document length (e.g. Feldman, Hazekamp, and Chawla (2016)); moreover, frequencies of co-occurrence serve to identify the dominant word combinations and formalize them in semantic patterns (e.g. Peterson and Liu (2018)). However, studies based on sublanguage theory typically do not differentiate between individual types of variants and how they can be related to constraints governing the system. While the theory does acknowledge the fact that sublanguages are shaped by communities of practice (Kittredge 2003), most studies do not pay respect to sociolinguistic variables and pragmatic factors; typically, sublanguages are analyzed as abstract systems, which are disassociated from the circumstances of production (e.g. Friedman, Kra, and Rzhetsky (2002)).

The case study presented here aims to combine both levels of analysis by first describing the characteristic variation patterns of individual sublanguages, and then linking them to those properties that characterize the subsystem in its entirety. This way, terminological preferences can be integrated into the characterization of clinical sublanguages.

4.2 Main Hypothesis

The central assumption underlying this study is that the choice of a term variant depends on semantic properties of the underlying concept, as well as

context factors. Such context factors can be of textual and extra-textual nature, such as adjacent words and the communicative situation. Since sublanguages differ with regard to their semantic structure, as well as stylistic and pragmatic properties, it is likely that they also differ systematically with regard to term choices. These differences should manifest themselves in the distribution of domain-specific term types (i.e. terms showing particular features in the surface form) across sublanguages. For instance, it is likely that some sublanguages employ a higher proportion of term abbreviations than others, depending on the communicative context. Consequently, the hypothesis is that sublanguage features can be used to predict the occurrence of term types in clinical writing.

4.3 Outline of the Analytical and Experimental Procedure

To validate the hypothesis formulated above, the second part of this thesis presents a case study. This case study analyzes a sample of EHRs from a clinical specialty, namely endocrinology. The individual EHRs are composed of different sections relating to different aspects of a clinical encounter. Hence, it is argued that, while all EHRs are concerned with similar clinical cases, they employ a set of distinctive sublanguages. In Chapter 5, the dataset under analysis is introduced. The individual sections are characterized with regard to their function within the EHR, their thematic focus, as well as stylistic properties.

For a quantitative analysis of the semantic structure of these sublanguages, all EHRs are annotated with concept IDs from a medical terminology, SNOMED CT. Based on the annotated terms, the overall proportion of semantic types in the sample, as well as their distribution across the individual sublanguages, is calculated. Chapter 6 outlines the annotation procedure, discusses some of the methodological challenges encountered in the process, and presents the output of the task.

To compare term preferences across sublanguages, the terms need to be classified by formal criteria. Based on a subset of the identified terms, a feature set reflecting variation patterns is developed. This feature set is used to encode the formal properties of all identified terms, and to quantify the proportion of term types. In Chapter 7, the development of the feature set, the formal annotation and the output of this task are discussed in detail.

In the third part of this thesis, statistical modeling techniques are used to validate sublanguage-specific variation patterns. Focusing on a small concept sample, a number of classification experiments are conducted. The goal of these experiments is to evaluate the strength of the association between different types of context factors and formal properties of the terms. Chapter 8 describes the experimental setup and summarizes the results.

Chapter 5: The Clinical Dataset

This chapter introduces the dataset that is at the core of the case study. The first section provides some general information on the origin and size of the EHR sample, as well as the general structure of the documents (Section 5.1). The second section describes the individual sections of the documents in more detail, exposing their semantic and stylistic characteristics (Section 5.2). Based on this description, the final section of this chapter characterizes the languages used in the individual EHR sections as distinct sublanguages (Section 5.3).

5.1 Overview of the Dataset

The EHR sample was provided by the department of endocrinology at UZ Leuven. A set of 499 cases of diabetes was chosen from the pool of patients treated at this department. The patients visit the hospital for regular check-ups, typically every 3 months. For each patient, all associated EHRs were retrieved from the clinical data warehouse, and exported in plain text format. In total, the dataset comprises 13,359 EHRs (i.e. documents relating to individual consultations), which were composed by 636 clinicians between 1998 and 2016. The total length of the dataset is 3,669,097 tokens. On average, the case history of one patient is 7,353 tokens long and contains 27 individual EHRs. The mean length of an individual EHR is 273 tokens.

All EHRs were de-identified by the ICT department. The de-identification procedure involved the removal of PHI relating to the patient as well as the clinicians involved in their treatment. Within the text files, all names, places and contact information were replaced by placeholders (e.g. @name@, @date@). In addition, the names of the clinicians who composed the EHRs were replaced by numerical IDs.

5.2 Structure of the EHRs

All EHRs were composed with a semi-structured template, which contains different sections covering the different stages of a clinical encounter. Typically, a consultation starts with a summary of the patient's medical history. Following the verbal assessment of the patient's state, a physical examination is carried out. Based on the findings of the consultation, a medical conclusion is formulated, which states new diagnoses or confirms existing ones, and outlines the further course of therapy. After the consultation, selected sections of the EHR are exported and merged into a letter, which is forwarded to the patient's GP.

The EHR template has some fields that contain merely numerical or binary values (e.g. concerning the weight or smoking status of the patient). However, as these items provide little insight for terminological analysis, they are not included in the study. After the exclusion of such pre-structured elements, the dataset still features 41 different free-text sections. However, not all of them appear in every EHR: The template used for data entry has evolved over the years, reflecting changes in clinical practice and documentation. For example, the more recent EHRs contain a dedicated section relating to the frequency and severity of hypoglycemic events; in the older EHRs, this information is found in different parts of the document. Some EHRs also contain reports on special procedures, which are not routinely conducted at every consultation, and, therefore, only appear in a small number of EHRs (e.g. the report on impedance analysis, which only occurs in 10 documents). Moreover, there are some administrative artifacts caused by the conflation of data from different departments (e.g. *Electrocardiography Report 1* and *Electrocardiography Report 2*, both of which obviously relate to identical procedures).

For a detailed characterization of sublanguage features, the focus of this study is on the ten most frequent sections, i.e. those that appear in the highest number of EHRs. Still, among the top sections, there are major differences in frequency: Those sections that cover the core elements of a consultation, such as the *Conclusion*, the patients' *Complaints*, and the medical *History*, appear

in more than 90% of the EHRs; by contrast, the *Eye Report*, which is a specialist report summarizing ophthalmological investigations, only occurs in about a third of the documents. Moreover, the length of the sections differs considerably: The average length of the *Conclusion* is 69.07 tokens (accounting for 25.30% of the total number of tokens present in an average EHR), while the average length of the *Diet* section is only 8.84 tokens (corresponding to 3.24%).[3] The general statistics of the core sections are provided in Table 1. In the following, the sublanguages used in the core sections are characterized in more detail.

Table 1: Overview of the core EHR sections, sorted by their total frequency in the dataset (i.e. the number of EHRs in which they appear). The second and third column specify the total and the relative frequency of each section (i.e. the proportion of EHRs containing this section relative to the total number of EHRs in the dataset). The fourth column gives the average length of the section in tokens. The fifth column specifies the relative length of the section (i.e. the ratio of the average number of tokens in this section and the total number of tokens in an average EHR).

Section	Total frequency of the section	Relative frequency of the section in %	Average length in tokens	Relative length in %
Conclusion	13,162	98.53	69.07	25.30
Complaints	12,580	94.17	38.02	13.93
History	12,384	92.70	62.35	22.84
Diet	11,633	87.08	8.84	3.24
Comments	11,331	84.82	17.68	6.48
Examination	10,295	77.06	15.53	5.69
Anamnesis	9,860	73.81	20.80	7.62
Medication	8,190	61.31	28.98	10.62
Therapy	7,786	58.28	11.98	4.39
Eye Report	4,296	32.16	17.81	6.53

5.2.1 Anamnesis

The *Anamnesis* serves to assess the environmental and behavioral factors that affect the patient's condition. Therefore, this section covers a very wide

[3] Unless specified otherwise, all average values are calculated by the arithmetic mean.

semantic spectrum, including a high proportion of non-medical details. For example, the *Anamnesis* can provide information on lifestyle factors (e.g. physical exercise and alcohol intake), genetic risk factors (e.g. prevalence of particular diseases among family members), the family situation or the employment status of the patient. As illustrated by Examples (1) and (2), these factors can be interspersed with details on the current therapeutic regimen:[4]

(1) *Hij heeft een normale intake. Spuit onmiddellijk na de maaltijd. In de vakantie Lantus naar omhoog getrokken (van 24E (eenheden) naar 28 E). Vooral omdat hij anders naar de ochtend toe te hoog stond, zeker omdat hij in de vakantie overdag weinig actief is.*

'He has a normal intake. Injects immediately after meals. Increased Lantus during holidays (from 24 units to 28 units). Mostly because otherwise his values were too high in the morning, probably because during the holidays he is not very active during the day.'[5]

(2) *@name@, werkt voor @name@ echtgenote verblijft meestal in @place@ Hij doet geen sport. Wandelt veel. Familiaal: geen diabetes. gaat 3x per week voor werk op restaurant en reist veel dochter is arts*

'@name@ works for @name@ wife usually stays in @place@ He does not do sports. Walks a lot. Familial: no diabetes. goes to restaurant 3x per week for work and travels a lot daughter is a doctor'

[4] In accordance with the privacy agreement with UZ Leuven, no part of the data can be cited directly. Therefore, all examples cited in this section are fictitious. To illustrate the particular linguistic features of the individual sections, syntactic structures and words were borrowed from the original text. However, to prevent the association with individual cases, they have been modified by mixing elements from different EHRs, changing the sentence order or replacing individual words.

[5] All examples were translated as closely as possible. Reduced words were left in their original form in the Dutch example, with expansions provided in brackets. Misspellings were not corrected in the Dutch examples. In the English translation, the full forms are used for better readability.

As this section is composed in interaction with the patient, it is comparatively narrative in style. While many sentences lack a grammatical subject, the proportion of verb phrases is high, especially in the description of general habits of the patient. Given the semantic structure of this section, non-specialized terms, which are typically spelled out in their full form, account for a major part of the vocabulary. Conversely, if abbreviations occur, they usually relate to medical details:

(3) *hypo (hypoglycemia) elk maal na volleybal stopt (insuline) pomp 15 min ervoor en herstart 30 min erna. BD (bloeddruk) gem (gemiddeld): 110/80*

'hypoglycemia every time after volleyball then stops insulin pump and restarts 30 min afterwards. blood pressure average: 110/80'

(4) *woont bij moeder thuisvpl (thuisverpleging) 1x per dag Gesproken over islet tx (transplantatie) – bg (bloedgroep) B, maar schrik, wil niet*

'lives with mother home care 1x a day Discussed islet transplantation blood group B, but scared, doesn't want to'

5.2.2 Comments

The *Comments* serve for the internal exchange among colleagues about the case, for instance to note the procedures to be scheduled. As they are not included in official documents, they tend to be written in a telegraphic style and employ a high proportion of reduced forms and jargon terms. As shown by the examples below, they can be extremely terse, consisting of mere sentence fragments of dense enumerations of jargon terms:

(5) *R/ (rest) idem advies @name@ volg (volgende) x (keer) ur (urinecollectie), oft (oftalmologisch nazicht)!*

'Rest idem advice @name@ next time urine collection, ophthalmological check-up'

(6) *glc (glucosemeting), hbA1c (hemoglobin-A1c-meting), lip (lipidenprofiel), sk (schildklieronderzoek), urine (urinecollectie) en oftalmo (oftalmologisch nazicht) volg x (volgende keer), co (controle) 3 ma (maanden)*

'glucose level measurement, hemoglobin A1c measurement, lipid profile, thyroid panel, urine collection and ophthalmological check-up next time, control in 3 months'

This section is also used to point out shortcomings in current treatment, such as the miscoordination of medication, or failure to refer a patient to another specialist:

(7) *stop Glucophage (waarom krijgt hij dit??)!!!!! Bellen met @name@*

'stop Glucophage (why does he get this??)!!!!! Call @name@'

(8) *lipiden (lipidenprofiel), iono (ionogram) en creat (creatininemeting), urine (urinecollectie), hoe zit dat nu met psycholoog????*

'lipid profile, ionogram and creatinine measurement, urine collection, what about the psychologist????'

Besides, the *Comments* provide room for the off-the-record evaluation of the patient's coping with his disease and compliance with therapy:

(9) *HOPELOOS: zegt letterlijk levenskwaliteit te zoeken in het zoete eten en wenst dit niet te veranderen want gaat toch niet lang leven waarden zeer stabiel en matig verhoogd, ? fictieve waarden (meter (glucometer) nooit mee maar wel waarden in (diabetes) boekje)*

'HOPELESS: says literally to seek quality of life in sweet food and would not like to change this because will not live long anyway measurements very stable and moderately elevated, ? fictitious values (never brings the glucometer but measurements in diabetes journal)'

Similar to interactions on social media platforms, clinicians also make non-standard use of orthography and punctuation (cf. also Examples (7) and (8)), or insert emotional interjections to emphasize their point:

(10) zucht- meten aub.....

'sigh- measure please.....'

(11) MMMMMMMMMMMETEN!!!!!!!

'MMMMMMMMMMMEASURE!!!!!!!'

5.2.3 Complaints

The *Complaints* section summarizes the patient's physical and mental state at the point of the encounter. All cases deal with a chronic disease, which is highly sensitive to the patient's lifestyle, and whose successful treatment depends on the patient's active collaboration. Therefore, this section covers a wide spectrum of medical findings. The observations made in this section range from physical symptoms, both related and unrelated to diabetes, over the patient's psychological state to his performance in self-therapy (e.g. the self-administration of medication):

(12) De patiënt komt binnen en zegt onmiddellijk zich niet goed te voelen, grijpt naar de borst. Deze morgen na opstaan wat owel, kort van adem doch niet echt retrosternale pijn. Hij is fors dyspneïsch bij minste inspanning. Blijkbaar toch reeds lichte achteruitgan sinds enkele weken. Recent nazicht op dagzaal geriatrie met echo (echografie) cor (pulmonale hypertensie). Geen hoest of fluimen. Glycemies behoorlijk. zeldzame hypo (hypoglycemia).

'The patient comes in and says immediately that he does not feel well, grabs at his chest. This morning slightly unwell after getting up, short of breath but no real retrosternal pain. He gets severely dyspneic at the least exercise. Apparently slight decrease for a couple of weeks already. Recent check-up at geriatric outpatient clinic

with echography heart (pulmonary hypertension). No cough or phlegm. Decent glucose levels. rarely hypoglycemia.'

(13) *De kracht in het Re (rechter) been neemt af. Hij kan niet meer stappen. Ook meer pijn in de li (linker) knie. Uw patient had zijn diabetesdagboek niet meegebracht. Er zijn minstens 2 maal per week ernstige hypos (hypoglycemias). De familie dient dan glucagon toe.*

'The force in the right leg is decreasing. He cannot walk anymore. Also more pain in the left knee. Your patient had not brought his diabetes journal. There are serious hypoglycemias at least twice a week. Then the family administers glucagon.'

Similar to the *Anamnesis*, the *Complaints* are based on the direct interaction with the patient, resulting in a narrative style. As this section reflects the patient's perspective, it contains a high proportion of lay terms and colloquial expressions. Especially in the subjective evaluation of the patient's state, vague paraphrases prevail. Conversely, findings based on systematic medical assessment, such as the measurement of the blood glucose level, tend to be expressed in concise specialized terms or abbreviations, even if they were carried out by the patient themselves:

(14) *Schomelende glycemies. Alleenwonend. Eet voor het slapengaan nog altijd 1 boterham. Heeft angst voor hypo's (hypoglycemies). Hypo's worden gevoeld. Weinig voorkomend. Is nerveus. Kan niet van plaats. Veel last van zijn linker been. Denkt dat gewicht stabiel gebleven is.*

'Fluctuating glucose levels. Living alone. Always eats 1 sandwich before going to sleep. Is afraid of hypoglycemia. Hypoglycemia is perceived. Occurring rarely. Is nervous. Cannot budge. Bothered by his right leg a lot. Thinks that weight remained stable.'

(15) *Nog steeds veel pijn ikv (in kader van) CRPS (Complex Regional Pain Syndrome), doch recent opstarten van Trileptal. Is op de sukkel*

zegt ze zelf, ziet het niet zitten. Neemt de Zocor niet, Atacand is vervangen door Aprovel. Er blijft een probleem van therapietrouw.

'Still a lot of pain in context of Complex Regional Pain Syndrome, but recently started on Trileptal. Is ailing as she says herself, cannot cope. Does not take the Zocor, Atacand was replaced by Aprovel. The problem with therapy compliance remains.'

5.2.4 Conclusion

The *Conclusion* provides a summary of the insights made during the consultation; therefore, it tends to be the longest part of the EHR. It lists the current findings, as well as the procedures conducted during the consultation and their outcome. Then, it formulates a medical diagnosis and gives recommendations for further treatment, which are implemented in collaboration with the patient's GP. Hence, while this section is semantically diverse, it primarily refers to medical concepts. Besides, the *Conclusion* is directly addressed to an external recipient; therefore, it is composed in a relatively well-formed style, employing full sentences and consistent punctuation:

(16) *Bij uw patiënte werd naar aanleiding van reactieve hypoglycemies een glucosetolerantietest verricht. Deze toont een gestoorde glucosetolerantie. Omwille van het ongewoon metabool profiel screenden we actief naar een onderliggende pathologie. Het betreft hier een vroegtijdige diagnose van diabetes type 1. Er werd op de raadpleging reeds empirisch gestart met metformine.*

'On account of reactive hypoglycemic events your patient underwent a glucose tolerance test. It revealed a disturbed glucose tolerance. Because of the unusual metabolic profile we actively screened for an underlying pathology. We are looking at an early diagnosis of diabetes type 1. Metformin was already empirically started up at the consultation.'

Usually, the *Conclusion* is formulated at the very end of the consultation. This section thus marks a change in information status, which is evident in the term choices. As the *Conclusion* states clinically confirmed findings and well-defined disorders, it is dominated by concise specialized terms, which appear either in their full form or as canonical abbreviations:

(17) *Uw patiënt werd verwezen omwille van een nieuwe diagnose van diabetes mellitus met momenteel een Hb (hemoglobine) A1c van 6.8%. Onze controle bevestigt negatieve antistoffen voor pancreas, insuline en GAD (glutamic acid decarboxylase), waardoor type 1 diabetes met grote zekerheid uitgesloten is. Een aanvullende bloedname werd verricht ter uitsluiting van monogenetische vormen van diabetes (MODY (Maturity Onset Diabetes of the Young)). Verder weerhouden wij en normaal TSH (thyreoïdstimulerend hormoon)-spiegel.*

‘Your patient was referred to us due to a recent diagnosis of diabetes mellitus with a current hemoglobin A1C of 6.8%. Our check-up confirms negative antibodies for pancreas, insulin and glutamic acid decarboxylase, whereby type 1 diabetes is excluded with high certainty. A supplementary blood sample was taken to exclude monogenetic forms of diabetes (Maturity Onset Diabetes of the Young). Furthermore, we note a normal level of thyroid-stimulating hormone.’

(18) *Uitstekende stabiele glycemieregeling, dankzij de persisterende endogene insulinesecretie. De aanwezigheid van betacelgerichte autoimmuniteit (GAD (glutamic acid decarboxylase)-As positief) duidt echter op type 1 diabetes (slow onset).*

Excellent stable regulation of blood glucose, due to the persisting endogenous insulin secretion. The presence of autoimmunity at the beta cells (glutamic acid decarboxylase antibodies positive) indicates type 1 diabetes (slow onset), though.

5.2.5 Diet

The *Diet* section describes the eating habits of the patient and his compliance with the dietary regimen. Besides, it reports on modifications of the medication scheme in response to meals:

(19) *Diabetesdieet neemt geen tussenmaaltijden eet wel iets voor slapen gaan*

'Diabetes diet does not eat snacks but does eat something before going to sleep'

(20) *Voeding wordt afgewogen. Bolust extra bij tussendoortje.*

'Food is weighed. Boluses extra with snacks.'

This section is typically short, composed in a telegraphic style and semantically extremely confined. As it mostly refers to food and temporal details, it is dominated by general language terms. Specialized terms are only used to relate to particular dietary schemes, medication or units of measurement. As the *Diet* section is mainly intended for internal documentation, it contains a high proportion of abbreviations, both for specialized and lay terms:

(21) *diabetes (diabetsdieet), AVVZ (arm aan verzadigde vetzuren) – ZA (zoutarm)*

'diabetes diet, low in saturated fats – low salt'

(22) *ontbijt: 3 BH (boterhammen) met fruit (5 KHRwaarden (koolhydraatruilwaarden)) middag: idem avond: warme maaltijd (5 KHRwaarden) geen tussenmaaltijden, enkel bij lagere waarden eet voor slapen (2KHRwaarden onder 150) uit angst voor hypo (hypoglycemie)*

'breakfast: sandwiches with fruit (5 carbohydrate exchange units) lunch: idem dinner: warm meal (5 carbohydrate exchange units) no

snacks, only when low values eats before going to sleep (2 carbohydrate exchange units below 150) out of fear of hypoglycemia'

5.2.6 Examination

This section summarizes the physical examination carried out during the consultation. Usually, it covers a fixed set of routine procedures, which are carried out to assess the general condition (e.g. the measurement of blood pressure) or to detect common complications of diabetes (e.g. diabetic foot). Therefore, the *Examination* section can be characterized by a limited vocabulary, referring to standard procedures, body parts under investigation, and the presence or absence of typical findings:

(23) *-0.9 kg Pols: 90/min. Voeten: pulsaties voelbaar, geen wonden. Longen zuiver VAG (vesiculair ademgeruis) bilateraal. Cor: geen geruisen. Geen perifere oedemen. Abdomen: soepel, normale peristaltiek. Schildklier palpabel. Geen palpabele adenopathieën.*

'-0.9 kg Pulse: 90/min. Feet: perceptible pulsations, no wounds Lungs clear Bilaterally vesicular breath sounds. Heart: no sounds. No peripheral edema. Abdomen: smooth, normal peristalsis. Thyroid palpable. No palpable adenopathies.'

The *Examination* mainly serves for the collection of evidence to support a medical diagnosis, i.e. for internal documentation. Similar to the *Comments*, it is composed in a telegraphic style and dense with clinical jargon. It contains few full sentences, but rather enumerations of conventionalized procedures, which tend to be expressed in abbreviated terms:

(24) *Pulm (Pulmonair): nl (normaal) VAG (vesiculair ademgeruis), geen crepitaties Cor: T1T2, ES (extra-systole), syst (systolisch) souffle 1/6 (moeilijk hoorbaar) Voeten: geen wondjes, pitting oedeem*

'Pulmonary: normal vesicular breath sounds, no crepitations Heart: T1T2 *(parameter relating to relaxation times),* extra systole, systolic heart murmur 1/6 (difficult to hear) Feet: no wounds, pitting edema'

(25) *licht oedeem OL (onderste ledematen), geen wondjes aan de voeten, CVD (centraal veneuze druk) lijkt nl. (normaal), reserve obesitas, nl. cortonen, nl. VAG (vesiculair ademgeruis) dp (drukpijn) linker flanks*

'light edema lower limbs, no wounds at the feet, central venous pressure seems normal, residual obesitas, normal heart sounds, normal vesicular breath sounds pressure pain left side'

5.2.7 Eye Report

The *Eye Report* documents the outcome of ophthalmological investigations, which are routinely conducted to detect visual impairments. In particular, it serves to identify diabetic retinopathy, which is a frequent complication. This section is thus a type of specialist report, which is composed at a different department and then forwarded to the treating clinicians at the department of endocrinology. It is very confined in scope and, typically describes a fixed set of procedures. Similar to the *Examination*, it can thus be characterized by a condensed style of writing; it mainly employs telegraphic constructions and a highly specialized terminology. In particular, the proportion of abbreviations is high:

(26) *visus re (rechts) en li (links) 1.0 geen tekens van DRP (diabetische retinopathie)*

'visual acuity right and left 1.0 no signs of diabetic retinopathy'

(27) *Ver OD (oculus dexter): 1.2, Ver OS (oculus sinister): 1.0, Lezen OD: Snellen (Snellenkaart) 1, Lezen OS: Snellen 1.*

Distant right eye: 1.2, Distant left eye: 1.0, Reading right eye: Snellen chart 1, Reading left eye: Snellen chart 1.

Besides, this section is used to state follow-up procedures, which are scheduled based on the current findings:

(28) *visus ODS (oculus dexter et sinister) 1.0 oogfundus: OD (oculus dexter): alles OK OS (oculus sinister): 1 puntvorminge bloeding nasaal van de macula of microaneurysma. Lichte diabetische retinopathie, maculopathie. Fluo (fluorescentie-angiografie) gepland*

'visual acuity right and left eye 1.0 eye fundus: right eye everything OK left eye: 1 dot-shaped bleeding in nasal position from macula or microaneurysm. Mild diabetic retinopathy, maculopathy. Fluorescein angiography scheduled'

5.2.8 History

This section summarizes the medical history of the patient by listing all known clinical events, both related and unrelated to diabetes, in chronological order. Similar to the *Conclusion*, the *History* tends to be a lengthy and semantically diverse section, which employs specialized terminology and canonical abbreviations to state confirmed diagnoses. However, stylistically, it differs considerably: In contrast to the *Conclusion*, it contains barely any full sentences, but merely enumerates prior conditions along with the date of their first occurrence:

(29) *Strabisme. Sick Sinus Syndroom met pauzes tot 3.7 sec waarvoor pacemaker (Follow-up @name@). @date@: Paresthesieën linkerhand en rechtervoet, neurologisch onderzoek negatief. @date@: liesbreukherstel. @date@: herstart Actrapid op proef. @date@: Inversietrauma van de linker enkel met avulsiefractuur van de mediale malleolus waarvoor gipsimmobilisatie. @date@: Ijlhoofdigheid vermoedelijk secundair aan orthostatisme door autonome neuropathie.*

'Strabismus. Sick sinus syndrome with pauses up to 3.7 sec wherefore pacemaker (follow-up @name@). @date@: paresthesias right hand and left foot, neurological examination negative. @date@: surgery for inguinal hernia. @date@: restart Actrapid on trial. @date@: inversion trauma of the right ankle with avulsion fracture

of the medial malleolus wherefore immobilization by plaster case. @date@: light-headedness presumably secondary to orthostasis caused by autonomic neuropathy.'

Apart from clinical events, the *History* specifies major changes in the course of therapy, such as the switch from one medication to another:

(30) *Vertigo. Gemengde hyperlipidemie. Hypotensie op Amlor. GI (gastro-intestinale) intolerantie voor Glucophage. @date@: switch naar Novorapid, Lantus*

'Vertigo. Mixed hyperlipidemia. Hypotension under Amlor. gastrointestinal intolerance for Glucophage. @date@: switch to Novorapid, Lantus'

(31) *@date@ en @date@: 2x insulinepomp wegens zwangerschap. Sectio met geboorte van een gezonde zoon op @date@ en van een gezonde dochter op @date@. @date@: sectio meisje @naam@ + tubaligatie postpartumbloeding en endometritis (pomptherapie tijdens zwangerschap). Insulineanalogen sinds @date@. Intolerantie metformine chronisch slechte glycemiecontrole sinds partus in @date@*

'@date@ and @date@: 2x insulin pump because of pregnancy. sectio (caesarea) with birth of a healthy son on @date@ and a healthy daughter on @date@. @date@: section girl @name@ + tubal ligation postpartum bleeding and endometritis (pump therapy during pregnancy). Insulin analog since @date@. Intolerance for metformin chronically poor glycemic control since partus in @date@'

Typically, this section is not composed from scratch at every consultation. Instead, clinicians tend to copy and paste the entire text from EHRs relating to previous consultations; updates are only made if required, e.g. to append a new diagnosis.

5.2.9 Medication

This section lists all the medication prescribed to the patient that is not directly related to the treatment of diabetes. Semantically, the *Medication* section is thus very homogeneous, relating exclusively to pharmaceutical products and details concerning the dosage or mode of administration. It mostly employs specialized terms, in particular names of active ingredients or commercial trade names, which are combined in dense enumerations:

(32) *Efexor 150 mg 1 – 1/2/d (dag) Seloken 100 mg 1x/d Vit (vitamine) D druppels om de 14 dagen Simvastatin 40 mg 1x/d Coversyl 5 mg/d*

'Efexor 150 mg 1 – 1/2/day Seloken 100 mg 1x/day Vitamin D drops every 14 days Simvastatin 40 mg 1x/day Coversyl 5 mg/day'

Abbreviations are commonly used to express units of measurements and details concerning the mode of administration. By contrast, the medication terms themselves are mostly spelled out in their full form. Reduced forms are only used for high-level pharmaceutical classes (e.g. *vit* 'vitamin' in Example (32) above) or active substances (e.g. *perindo.-amlodip.* 'perindopril amlodipine' in Example (33) below), but not for product names:

(33) *Asaflow tabl (tablets) 160 mg po (per os) 1.0 tabl – coveram (perindo.-amlodip.(perindopril-amlodipine)) tabl 10-5 mg po 1.0 tabl – emcoretic drag (dragées) mitis 5-12.5 mg po 1.0 tabl*

'Asaflow tablets 160 mg orally 1.0 tablet – coveram (perindopril-amlodipine) tablets 10-5mg orally 1.0 tablet – emcoretic mitis dragées 5-12.5 mg orally 1.0 tablet'

References to other types of entities are very rare. If they do occur, they typically specify the reason for a change in medication:

(34) *Aspegic 1000 mg 3x1 (sinds 2 weken in afbouwschema owv (omwille van) pericarditis)*

‘Aspegic 1000 mg 3x1 (in taper regimen for 2 weeks due to pericarditis)’

5.2.10 Therapy

This section summarizes the medication administered specifically for the treatment of diabetes. Similar to the *Medication* section, the *Therapy* mainly lists pharmaceutical products and gives advice on their administration. However, with regard to the pharmaceutical concepts, there is no overlap between the two: The *Therapy* refers to antidiabetica exclusively, while the *Medication* covers all other medication. Thus, the semantic scope of the *Therapy* is even more confined. At the formal level, the sublanguages used in the two sections are very similar. One particular feature of the *Therapy*, though, is the high proportion of temporal expressions. Since the timing of drug administration is crucial to the efficient treatment of diabetes, this section provides very detailed instructions, either in the form of absolute values, or relative to the patient’s daily rhythm and eating habits:

(35) *Basaal: 0.65 E (eenheden)/h van 08.00u tot 14u 1.0 E/h van 14.00u tot 22.00u 0.75E/h van 22.00u tot 00u*

‘Basal: 0.65units/h from 8 a.m. to 2 p.m. 1.0 unit/h from 2 p.m. to 10 p.m. 0.75units/h from 10 p.m. to 12 p.m.’

(36) *Vroege shift: Novorapid: 12 E (eenheden) – 8 E – 9 E SC (Subcutaan) Late shift: Novorapid: 18 E – 7 E – 10 E thuis: Novorapid: 17 E – 10 E – 9 E Bij nachtelijke hypo’s (hypoglycemies): Lantus: 46 E*

‘Early shift: Novorapid: 12 units – 8 units – 9 units subcutaneously Late shift: Novorapid: 18 units – 7 units – 10 units at home: Novorapid: 17 units – 10 unit – 9 unit In case of nightly hypoglycemia: Lantus 46 units’

In addition, this section can contain advice on the correct mode of administration, e.g. the anatomical location of insulin injections:

(37) *Toujeo 10 E (eenheden) ipv (in plaats van) Levemir Novorapid naar 7 E bij de warme maaltijd. Zones van lipodystrofie vermijden. Kortere naaldjes*

'Toujeo 10 units instead of Levemir Novorapid to 7 units with warm meals. Avoid lipodystrophic areas. Shorter needles'

As opposed to the *Medication* section, drug terms, including trade names, are routinely abbreviated. Reduced forms are common, especially to differentiate between different types of insulin:

(38) *AR (Actrapid) 26-9-9, Ins (Insuline) 30 E (eenheden)*

'Actrapid 26-9-9, Insulin 30 units'

(39) *AR (Actrapid) 12 E (eenheden) – 6 E – 12 E indien inspanningen, anders 14 E IT (Insulatard) 12 E*

'Actrapid 12 units – 6 units – 12 units if exercise, otherwise 14 units Insulatard 12 units'

(40) *7-12-12E (eenheden) NR (Novorapid)/ 40E LEV (Levemir) Losferon 1x/d (dag) Vit (Vitamine) B12 1x/3maanden*

'7-12-12 units Novorapid/ 40units E Levemir Losferron 1x/day Vitamin B12 1x/3months'

5.3 Sublanguage Differences between the EHR Sections

As illustrated by the previous section of this chapter, the different sections of the EHRs in our dataset differ considerably with regard to their linguistic properties. These differences manifest themselves at the semantic, syntactic and lexical level.

The *semantic composition* varies widely, depending on the thematic focus of the respective section. The spectrum reaches from sections devoted to the description of the general circumstances of life (e.g. *Anamnesis*), or a particular aspect thereof (e.g. *Diet*), over sections covering a wide range of concepts of medical nature (e.g. *History*), to those confined to a narrow clinical specialty (e.g. *Eye Report*), or a particular element of treatment (e.g. *Therapy*).

Likewise, the *syntactic complexity* differs between the sections, reflecting the respective circumstances of composition and their communicative function. Those sections based on verbal interaction (e.g. *Complaints*) have a narrative style, containing a high proportion of verb phrases and full sentences. Similarly, in the *Conclusion*, the need to communicate with external recipients in a comprehensible manner motivates the use of well-formed sentences. By contrast, in those sections that are mainly intended for internal documentation (e.g. *Examination*), a telegraphic style of writing prevails, which can be characterized by the omission of syntactic elements, such as verbs and function words; at the extreme end, we find highly nominalized sections, consisting of mere enumerations (e.g. *Therapy, Medication*).

Finally, the sections show a distinctive *lexical structure*. The proportion of general-language words and medical terms differs as a function of semantic specialization and syntactic complexity. With regard to the medical terms themselves, we see differences in terminological preferences, which can be attributed to pragmatic factors, including the knowledge level of the involved speakers and the communicative function: While the sections based on the direct interaction with the patient tend to use lay terms and vague paraphrases (e.g. *Complaints, Anamnesis*), those based on the clinicians' evaluation are dominated by concise, specialized terminology (e.g. *History*). On the other hand, the formality of the communicative context determines the degree of terminological standardization: While the *Conclusion*, which is directly addressed at an external recipient, is dominated by canonical terms, those sections intended for internal use (e.g. *Examination, Comments, Eye Report*) show a high proportion of reduced forms, including ad-hoc abbreviations.

Overall, the language found in the EHR sample shows typical sublanguage behavior. It can be characterized by the formal features introduced before (cf.

Section 3.1.2), including the *finiteness* of words and their combinations, the *skewed distribution* of words and word types and its *deviancy*, both at the syntactic and lexical level. However, the prominence of these features varies strongly across the different sections, which warrants their treatment as individual sublanguages. These sublanguages can be situated on a continuum: At one end, there are patient-centered, narrative sections, which might still be intelligible to lay people; at the other end, we find highly specialized sections dominated by jargon expressions, which might be incomprehensible even to medical experts from other clinical specialties.

These sublanguage properties influence the potential for different types of term variation: As detailed earlier (Section 2.2), *conceptual variation* can be caused by cognitive factors. In particular, the knowledge level and perspective of the experiencer influence the way they mentally classify a phenomenon, and how they verbally express it. As described above, the individual sections differ with regard to the included perspectives (i.e. lay vs. specialist); therefore, variation processes at the conceptual level manifest themselves between those sections that reflect the patient's experience, and those that are based solely on the clinician's judgment. Moreover, the information status changes in the course of a consultation, evolving from an unclassified phenomenon to a confirmed diagnosis. This can cause differences in term choices, in particular between those sections that precede the clinical examination, and those that are composed in retrospect. *Denominative variation* occurs mostly as a function of register. Across the sublanguages in our dataset, register differences manifest themselves primarily in two dimensions: Firstly, the degree of specialization varies between those sections that are based on doctor-patient interaction, and those that are intended for peer-to-peer communication. Secondly, the level of standardization differs between those sections that are included in the official communication with external readers, and those that serve the informal exchange between colleagues. These two factors can motivate variation processes at the denominative level, in particular alternations between vernacular vs. specialized terms, and standard vs. non-canonical variants. *Lexical variation* is mostly conditioned by rules of the linguistic system (i.e. grammar and orthography), and the degree to which these rules are implemented. As illustrated above, the sublanguages differ

with regard to their syntactic complexity and well-formedness, which determines their potential to show variation processes at the linguistic level.

Chapter 6: Annotation with Concepts from SNOMED CT

For a systematic description of term usage and variation in the dataset, a part of the EHR sample was manually annotated with concepts from a medical terminology. In total, 4,426 EHRs relating to 171 different patients were annotated; this corresponds to 33.13% of the EHRs and 34.27% of the patients in the dataset. The size of the annotated portion is 1,278,376 tokens, i.e. 34.84% of the total.

The annotation project involved two stages: In the first stage, the raw text was labeled with concept IDs from SNOMED CT; in the second stage, the term-concept associations were validated. This chapter gives an overview of the procedure and output of the annotation project: The first section describes the initial annotation stage (Section 6.1). After defining the formal aims, it describes the knowledge sources, the annotation tools and the procedure of the task. Following a similar structure, the second section outlines the aim, as well as the setup and procedure of the validation stage (Section 6.2). The third section presents the results (Section 6.3). First, it describes the primary output of the project, including the distribution of concepts, terms and semantic types (Section 6.3.1), and the ratio of terms and concepts (Section 6.3.2). Then, it evaluates the annotation project from a methodological viewpoint (Section 6.3.3). The final section (Section 6.4) summarizes the findings.

6.1 Annotation Stage

6.1.1 Aim of the Annotation Task

The aim of the annotation was to identify all medically relevant entities in free text and link them to the corresponding identifier in the clinical terminology SNOMED CT. To get a representative view on term usage in clinical

practice, the annotation task was designed to be as exhaustive as possible. There were no preliminary restrictions concerning the target entities, neither with regard to the semantic properties, nor the grammatical or formal features of the terms encountered in text. All entities that have a concept entry in SNOMED CT were considered as medically relevant, even if they do not pertain to a medical category in a narrow sense, but, for instance, general lifestyle. Regardless of the surface form, all occurrences of these concepts were annotated, including non-standard and lay terms.

6.1.2 Setup and Procedure

6.1.2.1 Knowledge Sources

As the primary knowledge source, SNOMED CT was used for the annotation with concept IDs. However, at this point, there is still no comprehensive release for Belgian Dutch available (cf. Section 1.4). Therefore, the UZ Leuven uses its own localized version for the coding of EHRs. Within the hospital network, this version can be accessed through a customized terminology browser. The in-house terminology is based on the International Release of SNOMED CT and updated in the same rhythm. The local version used for the annotation task was thus based on the most recent international release at the point of the annotation (i.e. the July 2017 International Edition (SNOMED International 2019)). In the in-house terminology, one Dutch term is available per concept. However, since the translation process is not carried out in a systematic manner, the proportion of translated concepts varies across semantic categories.

For the annotation of pharmaceutical entities, an additional knowledge source was required. Either version of SNOMED CT only provides concept entries for active substances and pharmaceutical classes, but does not list trade names of commercial products. Therefore, to assign a SNOMED CT code to a product name, the active ingredients of the product had to be identified in the first place. To this end, the drug compendium provided by the Belgisch Centrum voor Farmacotherapeutische Informatie (Belgian Center for Phar-

macotherapeutic Information (BCFI))[6] was consulted. This compendium lists all pharmaceutical products that are currently on the market in Belgium, along with their active ingredients and dosage information. It is freely available online and can be accessed through a web browser (Belgisch Centrum voor Farmacotherapeutische Informatie 2019).

6.1.2.2 Annotation Tool

The annotation was carried out in an editor developed by the ICT department of the UZ Leuven. This editor was linked to the in-house version of SNOMED CT. Previous studies have shown that automatic pre-annotations can speed up the annotation progress (Grouin and Névéol 2014; Roller et al. 2016). Therefore, the editor was enhanced with an auto-suggest function based on string match. For all forms that overlapped with a term in SNOMED CT, a list of suggested concepts was provided; then, the appropriate concept was selected and confirmed. The term base used for the auto-suggest function was updated in real time. As soon as a new term variant was identified, all other occurrences of this form were pre-annotated as well. For the remaining forms, the SNOMED CT concepts were assigned individually. If a text span was selected in the editor, it was automatically copied into the search window of the local terminology browser. The copied term was automatically matched against the term base. If a direct match was found, it could be directly linked to the term in text; otherwise, the terminology was searched by hand for variants of the term. The annotation of commercial drug names required an intermediate step: Firstly, the name was searched in the online browser of the BCFI; secondly, the active ingredients was pasted into the SNOMED CT browser to identify the underlying concept.

[6] The BCFI is a non-profit organization associated with the Belgian Federal Agency for Medicines and Health Products, which aims to provide independent information on pharmaceutical products. One of its main publications is the *Gecommentarieerd Geneesmiddelenrepertorium* (Commented Drug Compendium), which is updated every year and serves as a reference for medical practitioners (Belgisch Centrum voor Farmacotherapeutische Informatie 2018).

6.1.2.3 Annotation Procedure

The annotation was carried out by five Masters' students of biomedical sciences. The annotators received a set of guidelines describing the main principles of their task: They were instructed to identify all relevant entities and link them to the corresponding SNOMED CT concept according to the criteria defined above (cf. Section 6.1.1). They were encouraged to identify all types of variants, including misspellings (e.g. *isnuline (insuline)* 'insulin'), reduced forms (e.g. *nierinsuff (nierinsufficiëntie)* 'kidney insufficiency'), derivations (e.g. *glycemisch* 'glycemic') and paraphrases (e.g. *wazig zien* 'see vaguely'). They were supposed to select the most fine-grained concept available and mark up the longest contiguous text span relating to one concept. For instance, the term *NPH insuline* 'Neutral Protamine Hagedorn insulin' should be labeled as *Isophane insulin (66384003)*[7] rather than *Intermediate-acting insulin (68475005)*, which is situated at a higher level in the conceptual hierarchy of SNOMED CT. Finally, if they encountered an habitual combination of semantic types (e.g. a finding and a body part), they were advised to select a compound concept if available, rather than annotating the individual constituents. For example, *abdominale obesitas* 'abdominal obesity' should be coded as *Central obesity (248311001)*, rather than *Obesity (414916001)* and *Abdominal structure (113345001).*

To prepare the dataset, all EHRs relating to one clinical case (i.e. one patient) were merged into a single text file. Five cases were reserved for training; the remaining 479 cases were randomly assigned to the individual annotators. As the consistency between annotators was calculated at the validation, rather than the annotation stage, no case files were retained for double annotation. First, all annotators labeled the set of training files to familiarize themselves with the task and resolve potential difficulties. After they received individual feedback, they proceeded to annotate their personal set of files. While the overall aim was to annotate as many cases as possible within the limited duration of the project, the annotators were allowed to work at their own pace.

[7] For this and the following concept examples, the numbers in brackets specify the SCTID of the concept in SNOMED CT.

6.2 Validation Stage

6.2.1 Aim of the Validation Task

The primary purpose of the validation task was to verify the term-concept associations obtained in the initial annotation stage. A term-concept pair was considered correct if the term was an adequate expression of the concept without additional context information. By contrast, a pair should be judged as invalid if it contained a reduced form, which could only be interpreted in an appropriate context; if there was a mismatch between the level of granularity, such that the term referred to a more general or specific concept; or if there was simply no semantic relation. In addition, the task served to assess the domain pertinence. In particular, a pair should be rated as domain-specific if the concept belonged to the domain of endocrinology.

6.2.2 Setup and Procedure

6.2.2.1 Knowledge Sources

The knowledge sources consulted for validation were identical with those used in the initial annotation stage (i.e. SNOMED CT and the BCFI drug compendium; cf. Section 6.1.2.1). However, for practical reasons, it was not possible to let the annotators work on-site at the UZ Leuven. As they worked remotely, the annotators had no access to the in-house version of SNOMED CT, which had been used for the initial annotation. Instead, they used the freely available International Edition,[8] which can be accessed through a web browser.

6.2.2.2 Validation Tool

The validation was carried out in a simple spreadsheet, which contained a list of term-concept pairs sorted by concept ID. For each pair, the spreadsheet

[8] https://browser.ihtsdotools.org/?

provided two checkboxes, one to rate the correctness, and one to judge the domain-pertinence of the pair.

6.2.2.3 Validation Procedure

The validation task was carried out by three Masters' students of biomedical sciences; two of them had already helped with the initial annotation. As in the annotation stage, they received guidelines describing the task: They were instructed to proceed through a list of unique term-concept pairs one-by-one. For each pair, they were supposed to look up the concept ID in the SNOMED CT browser and indicate in the spreadsheet whether the pair was, firstly, correct and, secondly, domain-specific. For example, the pair *amsleronderzoek* 'amsler examination' – *Amsler chart assessment (252885006)* should be judged as correct, but not domain-specific, whereas the pair *tsh gesupprimeerd* 'Thyroid stimulating hormone suppressed' – *Thyroid stimulating hormone suppression therapy (704078008)* should be rated as both correct and domain-specific. Conversely, the pair *strekken* 'stretch' – *Hand stretching (305076008)* should be marked as incorrect, since the term is ambiguous out of context. Likewise, the pair *schildklierpalpatie* 'thyroid palpation' – *Diagnostic palpation (417215002)* should be marked as incorrect, as the term is more specific than the associated concept.

In the initial annotation stage, 16,151 unique terms had been identified and linked to SNOMED CT concepts (cf. Section 6.3.1.1 for the detailed presentation of the results). The unique terms were split into 18 lists: One short list of 182 terms, corresponding roughly to 1/10 of the identified terms, was set aside to calculate the Inter-Annotator Agreement (IAA). The remaining terms were divided into lists of approximately equal length (800 – 1000 terms). As in the annotation stage, the annotators were allowed to work at their own pace. They pulled a term list from a shared directory and uploaded it upon completion, until all lists had been validated. In addition, each annotator validated the list of pairs set aside for IAA calculation and uploaded a personalized copy.

6.3 Results of the Annotation Project

This section presents the results of the annotation and validation tasks. First, it assesses the distribution of entities, unique terms and concepts across the EHR sections and discusses their domain pertinence and semantic structure (Section 6.3.1). Next, it quantifies the potential for variation among the annotated terms by calculating the concept-to-term and term-to-concept ratio (Section 6.3.2). Finally, it evaluates the methodology of the annotation project with regard to the scope of the annotated dataset, the consistency between annotators and common error sources during the annotation (Section 6.3.3).

6.3.1 Conceptual and Terminological Structure of the EHR Sections

6.3.1.1 Distribution of Entities, Unique Terms and Concepts

In the course of the initial annotation stage, 171 case histories were completely annotated. In total, 300,693 entities were identified. These were expressed in 16,151 unique terms, relating to 8,002 different concepts in SNOMED CT. After filtering out those terms that had been judged as incorrect in the validation stage, 274,082 entities remained. These entities correspond to 15,025 unique terms and 7,687 different concepts.

The distribution of medical entities is highly skewed across the sections: The vast majority, namely 271,176 entities (98.94%), occurred in the core sections described above (cf. Section 5.2.1 – 5.2.10). By far the largest share was identified in the sections relating to the clinical *Conclusion* (73.54%), followed by the medical *History* (13.88%) and the patients' *Complaints* (3.14%). However, the highest number of unique terms and concepts was found in the *Complaints*, closely followed by the *Conclusion* and *History*. Given the large discrepancy in the number of entities identified in these sections, this is a remarkable result. Compared to the other smaller sections, such as the *Examination*, the *Complaints* thus show a very high degree of conceptual diversity, as well as an extreme potential for term variation. Table

2 gives an overview of the distribution of the annotated entities across the core EHR sections, as well as the number of unique concepts and terms occurring in these sections.

Table 2: Distribution of annotated entities across the core EHR sections. The first column specifies the name of the section. The second and third column provide the values for the absolute number of entities identified in the respective section, and their proportion relative to the total number of entities annotated in the entire dataset (i.e. 274,082). For columns 2 and 3, the sum is given in the last row. The fourth and fifth columns specify the number of unique terms and concepts identified in the respective section. In the last two columns, terms and concepts occurring in more than one section are counted multiple times, i.e. once in every section where they occur.

Section	Number of entities	Proportion of entities in %	Number of unique terms	Number of unique concepts
Anamnesis	3,251	1.19	637	537
Comments	3,317	1.21	1,938	1,278
Complaints	8,607	3.14	5,687	3,063
Conclusion	201,562	73.54	4,737	3,056
Diet	1,395	0.51	173	119
Examination	5,050	1.84	1,745	1,009
Eye Report	1,888	0.69	1,027	546
History	38,044	13.88	2,408	1,932
Therapy	1,229	0.45	636	463
Medication	6,833	2.49	1,146	717
Sum	271,176	98.94		

6.3.1.2 Domain Pertinence and Semantic Structure

6.3.1.2.1 Distribution of General and Domain-Specific Entities across the EHR Sections

Most of the annotated entities relate to general medical concepts (217,151 entities, corresponding to 79.23% of all entities identified in the dataset). Only about one fifth expresses concepts that are pertinent to the domain of endocrinology (56,931 entities, or 20.77%). Across the core EHR sections, the proportion of domain-specific entities varies strongly: With 37.67%, the *Therapy* has the highest proportion of domain-specific concepts, followed by the *Conclusion* (25.36%) and *Comments* (19.84%). The lowest values were

found in the *History* (7.12%), *Examination* (4.63%) and *Medication* (2.81%). Table 3 provides the full results.

The varying proportions of domain-specific entities clearly reflect differences in the thematic focus of the sections. For instance, both the *Therapy* and *Medication* deal mostly with drugs, their dosage and administration; however, while the *Medication* refers to all kinds of substances administered to the patient, the *Therapy* section serves specifically to document antidiabetic drugs, resulting in a very high proportion of domain-specific entities. The *History*, while being very dense in specialized terminology, summarizes events from the entire clinical spectrum; hence, the proportion of domain-specific terms is rather low. Likewise, the low value in the *Examination* can be attributed to the fact that this section covers a fixed set of routine procedures, which serve to assess the general health of the patient. Conversely, in the *Eye Report*, whose aim is to detect specific complications of diabetes, the proportion of domain-specific entities is higher.

Table 3: Proportion of domain-specific concepts among the entities identified in the core EHR sections.

Section	Proportion of domain-specific entities in %
Anamnesis	7.41
Comments	19.84
Complaints	12.40
Conclusion	25.36
Diet	14.62
Examination	4.63
Eye Report	16.63
History	7.12
Therapy	37.67
Medication	2.81

6.3.1.2.2 Proportion of Semantic Classes among the Annotated Entities

For the semantic analysis of the annotated entities, the semantic groups of the UMLS, rather than the original categories of SNOMED CT, were used. This choice was motivated by the fact that SNOMED CT uses an extremely fine-grained semantic categorization scheme. For example, the phenomenon of

sweating is represented by two distinct concepts, which belong to different semantic classes: *Sweating (finding) (415690000)* and *Sweating (observable entity) (364538006)*. As pointed out by earlier research, such distinctions may be justified by ontological design principles, but may be overspecified to encode term usage in clinical practice (Fung et al. 2005; He et al. 2012). Moreover, with an eye on the final step of the research project, namely the modeling of term variation (cf. Chapter 8), such fine-grained distinctions may introduce artificial boundaries, which could blur variation patterns. Conversely, the semantic network implemented in the UMLS enables the semantic analysis at a coarser level: Every UMLS concept is assigned a semantic type; in addition, every type is associated with a broader semantic group (McCray, Burgun, and Bodenreider 2001). For example, in the UMLS, both the types *Clinical Drug* and *Pharmacologic Substance* belong to the group CHEMICALS & DRUGS (cf. National Library of Medicine (2018) for the full list of semantic groups and associated types).

All concepts were automatically mapped to their equivalent in the latest UMLS release (National Library of Medicine 2019a). Then, the semantic group was inferred based on the semantic type tag attached to the concept.

In total, 15 semantic groups are present in the dataset. DISORDERS are the most frequent group, followed by PROCEDURES, CONCEPTS & IDEAS, CHEMICALS & DRUGS and ANATOMY. Together, the top five semantic groups account for 91.39% of all annotated entities. Table 4 provides the examples for the semantic types associated with each group, as well as the detailed figures of the absolute and relative frequency of each group among the annotated entities.

Table 4: Distribution of semantic groups among the annotated entities, sorted by frequency. The first column specifies the name of the semantic group in the UMLS, with the standard abbreviation in brackets. The second column gives examples for the semantic types belonging to this group. The third and fourth column provide the total number of entities belonging to this group, and their proportion relative to all entities that were annotated in the dataset. The last row provides the sums of these values.

Semantic group	Examples of the semantic types associated with this group	Number of entities	Proportion of entities in %
DISORDERS (DISO)	Disease or Syndrome Sign or Symptom Anatomical Abnormality	95,089	34.69
PROCEDURES (PROC)	Diagnostic Procedure Therapeutic or Preventive Procedure Educational Activity	66,576	24.29
CONCEPTS & IDEAS (CONC)	Qualitative Concept Temporal Concept	48,227	17.60
CHEMICALS & DRUGS (CHEM)	Clinical Drug Hormone	26,044	9.50
ANATOMY (ANAT)	Body Part, Organ, or Organ Component Body Substance	14,534	5.30
LIVING BEINGS (LIVB)	Family Group Professional or Occupational Group	7,065	2.58
PHYSIOLOGY (PHYS)	Physiologic Function Mental Process	6,856	2.50
PHENOMENA (PHEN)	Laboratory or Test Result Biologic Function	5,684	2.07
OBJECTS (OBJC)	Manufactured Object Food	1,588	0.58
ACTIVITIES & BEHAVIORS (ACTI)	Daily or Recreational Activity Individual Behavior	772	0.29
DEVICES (DEVI)	Drug Delivery Device Research Device	638	0.23
OCCUPATIONS (OCCU)	Biomedical Occupation or Discipline	571	0.21
ORGANIZATIONS (ORGA)	Health Care Related Organization Self-Help or Relief Organization	338	0.12
GEOGRAPHIC AREAS (GEOG)	Geographic Area	78	0.03
GENES & MOLECULAR SEQUENCES (GENE)	Gene or Genome Amino Acid Sequence	22	0.01
Sum		274,082	100

The proportion of semantic groups varies across the EHR sections (cf. Table 5), reflecting their function in clinical documentation. In 6 of the 10 top sections, DISORDERS are the most frequent type. In particular, in the *Examination*, *Complaints* and *Eye Report*, they make up more or close to half of the annotated entities. By contrast, PROCEDURES are most frequent in the *Comments* and *Diet*. As expected, in the *Medication* and *Therapy* sections, CHEMICALS & DRUGS are the dominant group.

Moreover, the distribution of semantic groups reveals variations in the degree of semantic homogeneity: The medication-centered sections, i.e. *Medication* and *Therapy*, are semantically most constrained. While the remaining sections are more heterogeneous, some patterns become evident in pairwise comparison: For example, the *Anamnesis* and *Complaints* are relatively similar regarding the proportion of DISORDERS and PROCEDURES; however, in the *Anamnesis*, which assesses circumstances of daily life, there is a higher proportion of entities pertaining to the groups of ACTIVITIES & BEHAVIORS and LIVING BEINGS. The *Examination* and *Eye Report* both report on physical examinations of the patient. Interestingly, in both sections, the majority of the entities refer to DISORDERS, rather than PROCEDURES. One reason might be that these sections document a fixed set of routine procedures, which serve to assess the general health and detect common complications. As the methods used for this purpose are obvious to any domain specialist, they need not be mentioned explicitly. Instead, in these sections, there is a tendency to only state the presence or absence of findings (i.e. DISORDERS). However, there are differences in the proportion of entities belonging to the groups of ANATOMY and CONCEPTS & IDEAS, which can be attributed to the anatomical scope of the sections: As the *Eye Report* only deals with one body site, there is no need to specify the anatomical location. Instead, relative spatial modifiers are used for clarification (e.g. *right* or *left* eye); therefore, the proportion of entities belonging to the group of CONCEPTS & IDEAS is relatively high. Conversely, the *Examination* describes the investigation of different body parts; therefore, explicit references to anatomical locations are required.

Table 5: Proportions of semantic groups among the entities identified in the individual EHR sections. The rows specify the UMLS semantic groups, and the columns the EHR sections. For each section, the value relating to the most frequent semantic group is set in bold.

	Anamnesis	*Comments*	*Complaints*	*Conclusion*	*Diet*	*Examination*	*Eye Report*	*History*	*Medication*	*Therapy*
ACTI	5.94	0.83	1.79	0.44	1.97	0.18	0.79	0.28	0.03	0.75
ANAT	2.21	5.29	5.08	3.59	1.08	15.48	8.72	11.61	0.46	1.51
CHEM	6.31	21.21	7.23	7.09	3.69	0.90	2.15	6.80	**83.11**	**62.42**
CONC	10.22	7.81	10.40	20.79	22.70	6.87	14.06	5.41	8.95	11.92
DEVI	0.93	0.23	0.73	0.17	0	0.39	0.39	1.26	0.25	0.83
DISO	**44.55**	19.20	**49.86**	**31.24**	23.97	**53.52**	**48.05**	**45.00**	1.85	9.28
GENE	0	0	0	0.01	0	0	0	0	0	0
GEOG	0.42	0	0.19	0	0.7	0	0	0.07	0	0
LIVB	11.52	3.08	2.78	4.97	0.89	0.49	1.14	0.88	0.29	0.91
OBJC	2.01	2.32	2.34	0.56	15.58	0.21	0.04	0.21	0.98	1.13
OCCU	0.45	1.11	0.44	0.26	0.06	0	0.74	0.20	0.01	0.30
ORGA	0.62	0.49	1.10	0.20	0	0.16	0.09	0.22	0.29	0.08
PHEN	0.40	0.83	1.39	2.52	0	1.76	0.22	0.91	0.01	0.15
PHYS	0.93	2.53	2.26	3.19	0.51	7.03	4.42	2.19	0.41	0.38
PROC	13.47	**35.06**	14.40	24.96	**28.86**	13.01	19.19	24.97	3.35	10.34

6.3.2 Concept-to-Term and Term-to-Concept Ratio

6.3.2.1 Concept-to-Term Ratio

The number of terms associated with one concept reflects the propensity for term proliferation. If conceptual properties interact with the potential for term variation, it is likely that the concepts associated with a high number of variants have some semantic features in common. Therefore, to investigate

whether certain semantic features pattern with term proliferation, the *concept-to-term ratio* was calculated.

On average, for each concept, 2.18 valid variants were annotated. For 2,804 concepts (36.48%), two or more terms were identified. However, the distribution of the variants across concepts is highly skewed: The majority of concepts is linked to only one term, while a small number of high-frequency concepts is associated with a high number of variants (cf. Figure 1).

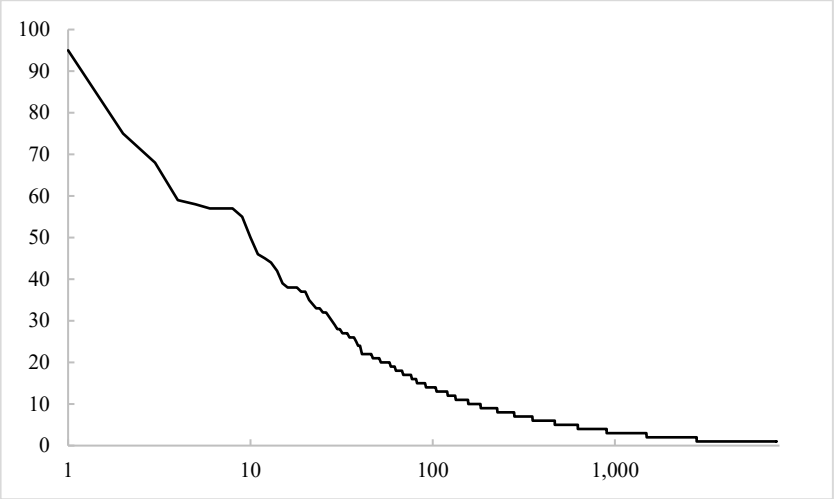

Figure 1: Concept-to-term ratio. The y-axis shows the number of associated variants. The x-axis shows the rank of the concept (i.e. the position if all concepts are sorted by the number of associated term variants in descending order) on a logarithmic scale.

The ten concepts with the highest numbers of associated term variants (cf. Table 6) are manually inspected. The majority, namely seven concepts, are instances of DISORDERS; the remaining three concepts belong to the groups of PROCEDURES, PHYSIOLOGY and CHEMICALS & DRUGS. Most of the concepts relate to findings that are either subjectively evaluated by the patient (*Patient feels well, Awareness*), or to measurements, substances and symptoms that are handled and recorded by the patients themselves, or with their collabora-

tion (*Change in insulin dose, Blood pressure recorded by patient at home*). This indicates that the inclusion of multiple perspectives can cause term proliferation. Besides, there are two concepts relating to findings made in the clinical setting (*Chest auscultation finding, Non-proliferative diabetic retinopathy*). What they have in common, though, is that they typically occur in informal sections documenting routine acts of investigation (*Examination, Eye Report*). As the use of jargon is acceptable in these contexts, the potential for variation, especially by means of term reduction, increases.

Table 6: Concepts with the highest numbers of associated terms. The first three columns specify the SCTID, the PT of the concept in SNOMED CT, and the semantic group of the concept in the UMLS. The last two columns provide the number of unique term variants, and the absolute frequency of the concept (i.e. the number of entities associated with this concept).

SCTID	PT	Semantic group	Number of term variants	Concept frequency
267112005	*Patient feels well*	DISO	95	139
446047003	*Change in insulin dose*	PROC	75	479
135815002	*General health good*	DISO	68	2686
301272007	*Chest auscultation finding*	DISO	59	201
262286000	*Weight gain*	DISO	58	275
312012004	*Awareness*	PHYS	57	157
39487003	*Insulin*	CHEM	57	1663
413153004	*Blood pressure recorded by patient at home*	DISO	57	132
390834004	*Non-proliferative diabetic retinopathy*	DISO	55	333
444780001	*High glucose level in blood*	DISO	50	1483

6.3.2.2 Term-to-Concept Ratio

On the other hand, terms that were linked to multiple concepts are likely to be polysemous, or instantiate a form of conceptual variation. Conceivably, terms sharing certain formal properties, such as abbreviations, are more prone to conflicting interpretations than others. Likewise, it is possible that certain semantic classes are more prone to show conceptual variation than others. To assess whether this is the case, the *term-to-concept ratio* was calculated.

On average, each term was linked to 1.12 concepts. 1,338 terms (8.91% of the unique terms) were associated with more than one concept; 31 terms (0.21%) had been labeled with five or even more concept codes. Figure 2 shows the number of concept associations across the unique terms.

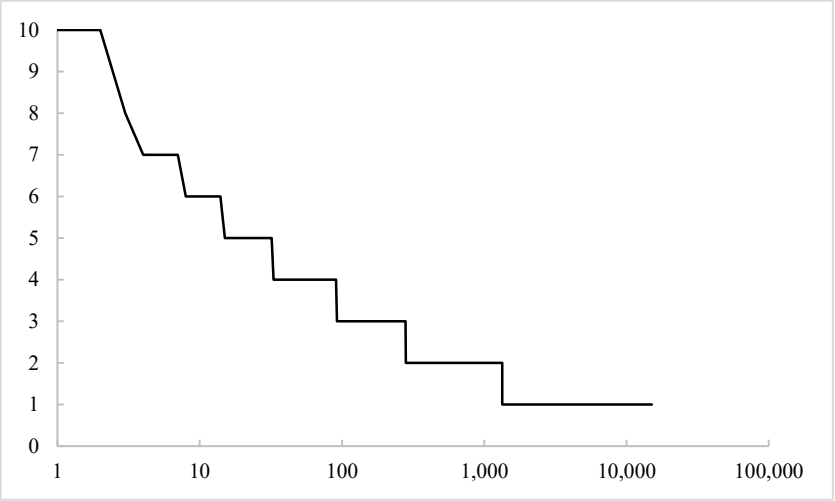

Figure 2:Term-to-concept ratio. The y-axis shows the number of associated variants. The x-axis shows the rank of the term (i.e. the position if all terms are sorted by the number of associated concepts in descending order) on a logarithmic scale.

The ten terms with the highest numbers of linked concepts (cf. Table 7) were reviewed individually. It is striking that all these terms tend to be used in the context of routine procedures and frequently occur in the informal sections of the EHR, such as the *Examination*. In all cases, the ambiguity arises from reduction processes, either at the linguistic or conceptual level, or the juxtaposition of both. On the one hand, we find abbreviations of the lexical form (e.g. *of*, which could be expanded to *oftalmologisch* ‘ophthalmologic’, *oftalmoloog* ‘ophthalmologist’ or *ophthalmologie* ‘ophthalmology’).

Table 7: Terms with the highest numbers of associated concepts. The first three columns specify the original term as annotated in text, the number of linked SNOMED CT concepts, and the absolute frequency of the term in the annotated part of the dataset. The last two columns give examples of the associated SNOMED CT concepts with their PTs and semantic groups.

Original term	Number of concepts	Term frequency	PTs of example concepts	Semantic group of example concepts
cor 'heart'	10	230	*Heart (80891009)*	ANAT
			Examination of heart (284448002)	PROC
abdomen	10	108	*Chest, abdomen, and pelvis (416775004)*	ANAT
			Procedure on abdomen (118698009)	PROC
pols 'pulse'	8	105	*Pulse rate finding (301147003)*	DISO
			Physiologic pulse (8499008)	PHYS
rr 'r-wave to r-wave interval'	7	7	*Finding of regularity of heart rhythm (301113001)*	DISO
			Normal heart rate (76863003)	DISO
creat 'creatinine'	7	354	*Creatinine (15373003)*	CHEM
			Measurement of creatinine clearance in peritoneal dialysis fluid specimen (442238003)	PROC
of 'ophthalmology/ ophthalmologist/ ophthalmologic'	7	63	*Ophthalmologic examination and evaluation (36228007)*	PROC
			Ophthalmology specialty (394594003)	OCCU
controle 'control'	6	4214	*Encounter for check up (185349003)*	PROC
			Self-control as a personality trait (284474008)	ACTI
urine	6	544	*Evaluation of urine specimen (442564008)*	PROC
			Urine – specimen type (122575003)	ANAT
visus	6	33	*Individual sight examination (171411003)*	PROC
			Eye/vision observable (363926002)	PHYS

On the other hand, lexical elements are omitted (e.g. *cor* 'heart', which can also be interpreted as a short form for 'examination of the heart'). In most cases, one of the associated concepts relates to a routine procedure, and one to the entity that is the object thereof (e.g. a body part or a physiologic function). Typically, the lexical element expressing the general act of exami-

nation is left out, since this semantic component can be inferred from context. In one case, polysemy is also caused by ambiguities at the morphological level: The clipped form *of* could be expanded to either a qualifier, a procedure described by this qualifier, a clinical specialty, or the person exerting this specialty. With one term, namely *rr*, the associated concepts differ merely in granularity, as they all relate to findings related to the heart rate. However, given the reduced form, the term has been linked to concepts at more or less specific levels: For instance, in the SNOMED CT hierarchy, *Normal heart rate* is a direct hyponym of *Finding of regularity of heart rhythm*.

6.3.3 Methodological Evaluation of the Annotation

6.3.3.1 Scope of the Annotated Dataset

One of the premises of sublanguage theory is that, since these languages deal with a confined subject matter, they only employ a limited set of terms (cf. Section 3.1). Thus, if a sample of texts in a specialized sublanguage is annotated, the rate at which new terms are encountered should decrease until a point of saturation (or *closure*) is reached, where all relevant terms have been acquired. As described earlier (cf. Section 3.2.1), previous research leveraged this phenomenon to determine the sublanguage status of a variety, or to measure the representativeness of a text sample with regard to a sublanguage. Here, closure properties are used to evaluate whether the scope of the project was sufficient to acquire an exhaustive terminology of the domain of endocrinology. If the annotated sample was large enough to be representative of the clinical specialty, the rate at which new concepts and terms had been acquired should show a decreasing trend.

To monitor the acquisition progress, the net rate of new concepts and terms acquired per case (i.e. all EHRs relating to an individual patient) was calculated in an iterative manner. Starting from the set of concepts and unique terms identified in the first case, the number of disjoint concepts and terms encountered in the next case was determined, and so on. To visualize the

global trend, the list of case histories (171 in total) was split into batches (17 batches of ten cases, and one batch of just one case). For each batch, the average acquisition rate was calculated, both for general medical and domain-specific concepts and terms.

On average, 3.19 domain-specific and 43.08 general new concepts had been identified in each batch. With regard to the unique terms, 10.36 domain-specific and 83.83 general terms had been acquired per batch. As is evident from Figure 3, though, the annotated EHR sample had neither been sufficient to reach a point of closure, nor to develop a decreasing trend at all. The rate at which new domain-specific concepts and terms were encountered is relatively low, especially considering the fact that the values are based on entire case histories, i.e. aggregated sets of EHRs. Presumably, this is an effect of the clinical domain under investigation: The dataset documents the treatment of a chronic disease, which progresses slowly over time, sometimes over decades. New diabetes-related diagnoses, or major changes in therapy are relatively rare, resulting in a low acquisition rate. Compared to that, changes in the general health or living circumstances are more frequent, and also more diverse across patients. This manifests itself in the higher acquisition rate and stronger fluctuations for general concepts and terms.

Hence, while it seems legitimate to consider the language used in endocrinology as a specialized sublanguage, the scope of the annotation project was clearly insufficient to obtain an exhaustive representation of the domain-specific terminology. This was already evident at an early stage of the annotation project, both due to cases of extreme term proliferation, and because the overall speed of the annotation was much below the expected level. However, due to limitations of time and resources, it was not possible to extend the project in order to annotate a larger part of the dataset.

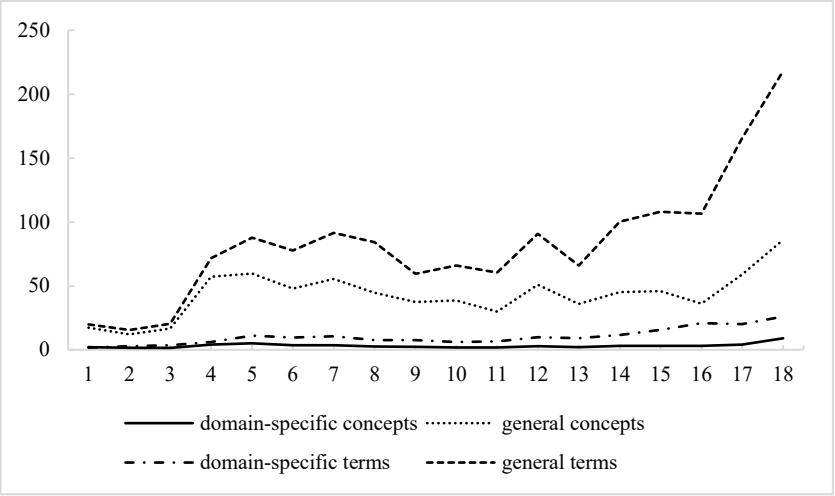

Figure 3: Progress of concept and term acquisition over the 171 annotated cases, averaged over 18 batches (17 batches of ten cases, and one batch containing just one case). The position on the x-axis indicates the batch number (i.e. the value at 1 refers to the average of case 1 to 10, the value at 2 refers to the average of case 11 to 20, and so on; the value at 18 specifies the values for the last case. The value on the y-axis specifies the average number of new concepts and terms acquired per case in the respective batch.

6.3.3.2 Inter-Annotator Agreement

To measure the reliability of the output generated by an annotation project, it is standard practice to have a part of the dataset labeled by all annotators, and calculate the consistency between their annotations (Artstein 2017). Given the slow progress in the initial annotation stage, and the fact that the two-stage design already implied a measure of quality assurance for the raw annotations, it was decided to calculate the IAA at the validation, rather than the annotation stage.

During the validation stage, one list of unique terms was set aside for IAA calculation and validated by all annotators (cf. Section 6.2.2.3). The IAA, as calculated by Fleiss' Kappa, was 0.51 for the validation of the correctness of term-concept associations, which is considered moderate. For the rating of

domain pertinence, the agreement was substantial with 0.77 (Landis and Koch 1977).

One reason for the low agreement in the rating of correctness could be the ambitious scope of the project. As the initial annotation task was designed to be as inclusive as possible in order to get a representative view on term usage in practice, the output was extremely diverse. The wide range of variation processes among the annotated terms made the validation a very complex task. While the guidelines aimed to specify how to judge different types of variants, they evidently still left room for conflicting interpretations. As the comparison of the personal copies of the agreement list shows, especially the reduced forms were a major cause of inconsistencies, as their acceptability was judged differently by the annotators. Besides, given the complexity of the SNOMED CT hierarchy, selecting the right level of granularity seemed to be an issue, especially for terms relating to pharmaceutical products, whose annotation required an intermediate mapping step.

6.3.3.3 Analysis of the Invalid Term-Concept Pairs

For a closer analysis of the errors that had occurred during the initial annotation stage, the invalid term-concept pairs were analyzed separately. Among these pairs, there were 1,126 unique terms and 315 unique concepts. Some terms had been linked to a high number of concepts and vice versa, such that the sample included for error analysis contained 1,746 unique pairs in total.

To quantify the different types of misclassifications, all of the erroneous term-concept-pairs were manually annotated with one of the following labels: *wrong concept* (there is no semantic relationship between term and concept); *term too vague* (the term does not cover all semantic facets of the concept); and *term too specific* (the term describes an entity that is more fine-grained than the concept).

Among the invalid pairs, the most frequent error type was *term too vague* (68.27%), followed by *wrong concept* (23.77%). Errors of the type *term too specific* only account for a minor part of the misclassifications (8.82%). This distribution illustrates the context-sensitivity of term usage in clinical prac-

tice. Most of the terms classified as *too vague* lack a constituent that can be inferred from habitual usage situations. For example, terms expressing a mere finding had frequently been linked to concepts that also specify an anatomical location (e.g. *kloofje* 'crack in the skin', which had been annotated as *On examination – cracked skin of feet* (164392008)). By contrast, most of the terms judged as too specific had been linked to a parent concept (e.g. *uitlezen glucometer* 'read out glucometer', which had been annotated as *Procedure categorized by device involved (363691001)*). For some of these terms, SNOMED CT indeed provides no concept at the right level of granularity, which pinpoints potential gaps in the conceptual hierarchy.

6.4 Conclusion

For a detailed analysis of term usage, the clinical dataset was partially annotated with concept IDs from SNOMED CT. The distribution of the annotated entities, terms and concepts provides further evidence for sublanguage differences between the different EHR sections: The thematic focus and communicative function of a section is directly reflected in the distribution of semantic categories among the entities, ranging from very homogeneous sections, which are dominated by entities from a single semantic group, to more heterogeneous ones, which refer to a wide range of medical and non-medical concepts. Likewise, fluctuations in the proportion of domain-specific concepts reveal differences in the degree of domain pertinence. At the same time, the sections show differences with regard to their terminological richness. In general, the number of annotated entities as well as the degree of semantic heterogeneity co-determine the number of unique terms per section. However, this is not the only factor: As the comparison shows, sections that involve a switch away from the canonical specialized language, either to a lay register (e.g. *Complaints*) or to informal peer-to-peer communication (e.g. *Comments*), are also highly prone to show term proliferation.

The analysis of the unique concepts and terms demonstrates that, while term variation is ubiquitous across the semantic spectrum, certain categories have a particularly high potential for variation: As the concept-to-term ratio shows,

concepts that depend on the subjective evaluation of the observer, especially if they are reported by a non-specialist, pattern with a high number of associated surface forms. On the other hand, the term-to-concept ratio indicates that high-frequency specialized concepts are highly prone to polysemy: As they are frequently referred to in fixed contexts, semantic components are habitually omitted; this can give way to conflicting interpretations.

The methodological evaluation illustrates the difficulty of documenting term usage in clinical practice. Evidently, the annotators struggled with the task of mapping entities from the entire semantic spectrum to a very fine-grained conceptual hierarchy. In particular, the richness of variation processes present in the data posed a challenge. Since the formal criteria were kept fairly liberal, the assessment of term validity proved a difficult task. Moreover, the analysis of closure properties shows that the creation of an exhaustive representation of term variants, even for a confined medical field such as endocrinology, would require immense resources. However, the output of the project provided evidence that the processes underlying the formation and usage of term variants can be linked to semantic, cognitive and pragmatic factors. Therefore, the systematic analysis of variation *types*, rather than that of individual *instances*, seems to be a promising approach.

Chapter 7: Annotation with Formal Term Features

For a more systematic description of the variation types present in the dataset, the terms identified by annotation were annotated a second time, this time at the formal level. All the unique terms were labeled with features reflecting the formal properties of the surface form. The first two sections of this chapter introduce the set of formal features (Section 7.1) and describes the annotation procedure (Section 7.2). The following section (Section 7.3) quantifies the distribution of the formal features among the identified terms (Section 7.3.1), and across the major semantic groups (Section 7.3.2). To conclude, the final section summarizes the findings and discusses the methodology (Section 7.4).

7.1 Feature Set

The variants identified by annotation show a wealth of variation types. In many cases, processes operating at different levels are juxtaposed: For example, the synonyms *nierfunctiebeperking* 'kidney function limitation' and *renale insuff (insufficiëntie)* 'renal insufficiency' differ with regard to their morpho-syntactical structure (one of them is a compound, and the other a pre-modified noun phrase), their register (one consists completely of lexical elements in the native language, while the other contains neo-classical elements), and the presence of reduction processes (in one variant, all constituents are spelled out, whereas the other contains a clipped form). The high combinatorial potential of variation processes complicates the operationalization by means of formal features: Obviously, the development of a tag set that adequately describes every variant would result in the explosion of term types. Therefore, a set of binary features was used, which was assigned to each variant with either positive or negative value; thus, each term can be

described by its individual feature configuration (e.g. *abbreviation* or not, *specialized* or not, and so on).

In total, eleven features were used. These features can be divided into three main groups, reflecting a term's *register*, the presence of *reduction* processes and its *morpho-syntactical* properties. Three *additional* features served to code *eponyms*, *trade names* and arbitrary *misspellings*. Table 8 gives an overview of the entire feature set. The following sections describe these features in detail.

Table 8: Overview of the formal feature set. The first two columns specify the group and the name of the feature. For each feature, the last two columns provide an example variant that shows this feature, and one that does not. Note that, for *eponym*, no negative example can be given, since no non-eponymous term exists; this is the case for the example cited here, as well as all eponymous terms identified in the dataset.

Feature group	Feature	Positive example	Negative example
Register	*Standard*	*arteriële hypertensie* 'arterial hypertension'	*ateriële hypertensie* 'arterial hypertension'
	Specialized	*pneumonie* 'pneumonia'	*longontsteking* 'lung infection'
Reduction	*Abbreviation*	*AHT*	*arteriële hypertensie* 'arterial hypertension'
	Lexical reduction	*24 h urine*	*24 h urine collectie* '24 h urine collection'
	Grammatical reduction	*roodheid rechter oog* 'redness right eye'	*roodheid van het rechter oog* 'redness of the right eye'
Morpho-syntax	*Compound*	*longauscultatie* 'lung auscultation'	*auscultatie van de longen* 'auscultation of the lungs'
	Derivation	*echografisch* 'echographic'	*echografie* 'echography'
	Paraphrase	*gewicht wil niet zakken* 'weight would not decrease'	*geen gewichtsverlies* 'no loss of weight'
Additional	*Eponym*	*ziekte van Raynaud* 'Raynaud's disease'	--
	Trade name	*Toujeo*	*Insulin glargine*
	Misspelling	*Isnuline*	*Insuline* 'Insulin glargine'

7.1.1 Register Features

To describe a term's register, the features *standard* and *specialized* were used.

To be considered *standard*, a variant had to fulfill two criteria: Firstly, it must be well-formed, i.e. comply with the conventions of orthography and grammar. For example, *arteriële hypertensie* 'arterial hypertension' *(Hypertensive disorder, systemic arterial (38341003))*[9] would be judged as a *standard* term, whereas the misspelled variant *ateriële hypertensie* would not; the phrase *onderzoek van de schildklier* 'examination of the thyroid' *(Thyroid panel (35650009))* would be considered *standard*, while the form *onderzoek schildklier* 'examination thyroid', which is missing grammatical function words, would not. As an exception, spelling variants of neoclassical terms, which can be attributed to regular alternations (e.g. *glycemie* and *glycaemie* 'glycaemia' *(Finding of blood glucose level (365812005))*, *oftalmologisch* and *ophthalmologisch (Ophthalmologic (239005))*) were considered *standard* regardless of the chosen variant. In general, abbreviations were considered non-standard. However, if a form is listed in SNOMED CT (e.g. *BP* for *Blood pressure (75367002)*), it was considered canonical and thus coded as *standard*. Secondly, a variant must be semantically equivalent to the reference terms in SNOMED CT. For example, the variant *diabetesretinopathie* 'diabetes retinopathy' covers the essential semantic components of the assigned concept, *Retinopathy co-occurrent and due to diabetes mellitus (4855003)*. This is not the case with the variant *sensorimotorische polyneuropathie* 'sensomotoric polyneuropathy', which had been linked to *Diabetic distal sensorimotor polyneuropathy (230573007).* Compared to the reference term, the annotated variant lacks two lexical components, *diabetic* and *distal*, which are relevant to convey the concept's semantics, in this case the etiology and location of the disease; therefore, it was not coded as *standard*. Similarly, in terms relating to pharmaceutical substances, only variants that

[9] In this and the following examples, the term in quotation marks provides a literal translation. The term and number in brackets specify the PT and the SCTID of the concept in SNOMED CT.

convey all the dosage and administration details specified in the reference term were labeled as *standard*. For example, the variant *Aspirin 300mg* had been linked to *Product containing precisely aspirin 300 milligram/1 each conventional release oral tablet (329525004)*. Since the variant lacks the information that the substance is delivered by an oral tablet, it would not be considered standard.

The feature *specialized* was assigned to terms that contain non-native roots. For example, *pneumonie* (annotated as *Pneumonia (233604007)*) was labeled as *specialized*, whereas the synonym *longontsteking* 'lung inflammation' was not. Likewise, the feature is assigned to abbreviations based on non-native roots, such as *AHT (arteriële hypertensie)* 'arterial hypertension' *(Hypertensive disorder, systemic arterial (38341003))*. As is typical for the medical domain, the majority of non-native terms is of neoclassical origin. However, there are also a number of loans from modern foreign languages, especially from English (e.g. *MRI (Magnetic Resonance Imaging (113091000))*) and French (*souffle* 'breeze' *(Aortic murmur (308687000))*). Commercial trade names of pharmaceutical products were coded as *specialized*, too (cf. Section 7.1.4).

7.1.2 Reduction Features

The reduction features served to encode three types of processes that result in a shortening or compression of a base term, namely *abbreviation*, *lexical reduction* and *grammatical reduction*.

The *abbreviation* feature was assigned to shortened lexemes, such as initialisms, acronyms, clippings, contractions and combinations of such processes, e.g. *g (gewicht) idem* 'weight the same' *(Weight steady (271398006))*; *DRP (diabetische retinopathie)* 'diabetic retinopathy' *(Retinopathy co-occurrent and due to diabetes mellitus (4855003))*; *ins (insuline) (Insulin (67866001))*; *skfctie (schildklierfunctie)* 'thyroid function' *(Finding of thyroid function (302074003))*. The *abbreviation* feature was assigned to non-standard forms, such as the previous examples, as well as to canonical abbreviations, which are listed as valid synonyms in SNOMED CT.

The feature *lexical reduction* was used for variants that, compared to the reference terms in SNOMED CT, lack a lexical element that is crucial to convey the concept's semantics. For example, the variant *24 h urine* had been linked to the concept *Urine specimen collection, 24 hours (2475000)*. However, the annotated variant does not state the procedure itself, but only the mode and object thereof. Similarly, the term *estradiol*, which had been annotated as *Estradiol measurement (37538009)*, lacks an explicit reference to the procedure. Therefore, these variants would be considered *lexical reductions*.

By contrast, the feature *grammatical reduction* was assigned to terms that contain all the relevant components, but miss function words that would be required to produce a grammatically correct form. For example, the variant *roodheid rechter oog* 'redness right eye' is semantically equivalent to *Red eye (75705005)*, but lacks the function words to explicitly link the finding to its anatomical location. Similarly, the variant *kst hersenen + stam (kernspintomografie hersenen en hersenstam)* 'MRI brain and brain stem', contains all the lexical constituents of the standard term, *Magnetic resonance imaging of brain and brain stem (29567006)*. However, as it is grammatically ill-formed, it is labeled as a *grammatical reduction*. Moreover, due to their reduced structure, all terms coded with this feature were also considered non-standard.

7.1.3 Morpho-Syntactical Features

The morpho-syntactical features were used to encode regular linguistic alternations, namely *derivations* and *compounding*, as well as *paraphrases*.

The *derivation* feature was mostly used for adjectives based on a nominal reference term. For instance, the variant *electrocardiografisch negatief* 'electrocardiographically negative' *(Electrocardiogram normal (164854000))* contains an adjective based on a standard noun denoting the procedure, namely *electrocardiografie* 'electrocardiography'. Similarly, the verb *infiltreren* 'infiltrate', which is based on the noun *infiltratie*, and had been linked to *Infiltration (231287002)*, would be annotated as a *derivation*. Besides, this

feature was used to code inflected forms of derived verbs, such as *stenoserend* 'stenosizing', which is based on the noun *stenosis (Stenosis (415582006))*.

The *compound* feature was assigned to all terms that consist of multiple lexical constituents, either of native, foreign or mixed origin (e.g. *bloedhoogdruk* 'high blood pressure' *(Hypertensive disorder, systemic arterial (38341003))*; *backgroundretinopathie* 'background retinopathy' *(Mild nonproliferative retinopathy (312903003))*; *longauscultatie* 'lung auscultation' *(Auscultation of the lower respiratory tract (449264008))*). This feature was also assigned to compounds with abbreviated constituents (e.g. *skf-tonderzoek (schildklierfunctieonderzoek)* 'thyroid function examination' *(Thyroid panel (35650009))*). However, only those terms where all lexical constituents can act as independent words were considered *compounds*. By contrast, complex neoclassical terms, which consist entirely of Greek or Latin confixes (e.g. *hypothyroïdie (Hypothyroidism (40930008))*), were not labeled with this feature.

The *paraphrase* feature is assigned to all terms that do not employ the reference term or one of its synonyms, but rather circumscribe the concept. For instance, the expression *gaat minder goed* 'going less well' would be coded as a paraphrase for *General health deterioration (285384003)*. Likewise, *gewicht wil niet zakken* 'weight would not decrease' *(Failure to lose weight (83421005))* and *slaapt niet goed door* 'does not sleep through well' *(Difficulty sleeping (301345002))* would be labeled with this feature.

7.1.4 Additional Features

Three more features were used to code *eponyms*, *trade names* and *misspellings*.

The *eponym* feature was assigned to terms that contain personal or geographical names. Typically, such terms refer to the person who was the first to

describe a body part or disorder, or who developed a medical procedure (e.g. *langerhanscel (Langerhans' cell (76322003))*; *ziekte van Raynaud (Raynaud's disease (195295006))*; *Snellen (Snellen chart assessment (252973004))*). Besides, this feature was used to label terms that make reference to professional associations (e.g. *NYHA (New York Heart Association) Klasse* 'New York Heart Association class' *(Assessment using New York Heart Association Classification (762998009))*).

All variants containing either product or company names were labeled with the *trade name* feature. For the greatest part, these terms relate to pharmaceutical products (e.g. *Toujeo (Product containing insulin glargine (126212009))*; *Merck-Bisoprolol 5 mg tabl (Product containing precisely bisoprolol fumarate 5 milligram/1 each conventional release oral tablet (318590006))*). A number of variants also refer to medical devices, such as pre-filled injection pens (*Solostar (Insulin autoinjector (706161000))*).

Finally, the *misspelling* feature was used to code arbitrary typos, such as *isnuline (insuline) (Insulin (67866001))*. This feature is thus only used for variations that cannot be attributed to paradigmatic alternations, as they are found in neoclassical terms or transliterations. For example, for the concept *Rheumatism (396332003)*, both *reuma* and *rheuma* were considered correctly spelled variants.

7.2 Annotation Procedure

The annotation was carried out in a spreadsheet containing the list of unique term-concept pairs (15,025 in total) and checkboxes relating to the individual features. To simplify the comparison with the reference terms, the pairs were aligned with the English and Dutch standard terms from SNOMED CT. For each pair, the set of associated terms was retrieved from the most recent release files of the International and Dutch editions of SNOMED CT (International Health Terminology Standards Development Organisation (IHTSDO) 2018; National ICT Instituut in de Zorg (Nictiz) 2018).

The list of term-concept pairs was traversed one by one. For each variant, the feature values were assigned by comparison with the reference terms. In addition, for pairs relating to pharmaceutical substances, the BCFI drug compendium (cf. Section 6.1.2.1) was consulted to assign the term features.

7.3 Results

This section presents the results of the formal annotation. First, it describes the global distribution of the formal features across the unique terms (Section 7.3.1). Then, to shed more light on the potential interaction of cognitive and conceptual factors with variation processes, it discusses the proportion of formal types across the major semantic groups (Section 7.3.2).

7.3.1 Distribution of Formal Features across the Unique Terms

The distribution of the register features shows that almost two thirds of the terms (61.62%) contain *specialized* elements. By contrast, only 38.94% of the terms are rated as *standard*. Notably, this does not even imply that these terms are actually included in SNOMED CT, but only that they would fulfill the formal criteria; hence, the number of terms that are actually documented is even less than that. Table 9 provides the detailed results for all features.

With regard to reduction processes, *abbreviations* are most frequent (15.44%), followed by *lexical reductions* (11.99%) and *grammatical reductions* (3.47%). At the same time, the different reduction mechanisms can also co-occur in one form. These figures demonstrate the high prevalence and complexity of reduction processes in a professional environment where time efficiency is crucial, resulting in highly ambiguous forms. Especially for *abbreviations*, which can allow for multiple expansions, and *lexical reductions*, which do not explicitly mention the essential properties of the concept, the correct interpretation strongly depends on context and domain knowledge.

Table 9: Distribution of formal features across the unique terms. The first two columns specify the group and name of the feature. The last two columns provide the total number of terms labeled with this feature, and their proportion relative to the number of unique terms.

Feature group	Feature	Number of terms coded with this feature	Proportion of terms coded with this feature in %
Register	*Standard*	6,540	38.94
	Specialized	10,349	61.62
Reduction	*Abbreviation*	2,593	15.44
	Lexical reduction	2,014	11.99
	Grammatical reduction	582	3.47
Morpho-Syntax	*Compound*	3,908	23.27
	Derivation	597	3.55
	Paraphrase	1,839	10.95
Additional	*Eponym*	141	0.84
	Trade name	1,774	10.56
	Misspelling	758	4.51

Almost one quarter of the terms are *compounds*, which indicates a strong tendency toward nominalization, even for the expression of complex concepts. This is typical for specialized discourse in general, and Germanic languages, like Dutch, in particular (Bretschneider and Zillner 2015). Conversely, only 3.55% of the variants are *derivations*. This low figure can also be attributed to the fact that not all terms are suited for derivation processes: While neoclassical terms are morphologically flexible, many native terms as well as trade names are not. *Paraphrases*, on the other hand, are more frequent. One reason could be that the data is partially based on verbal interactions with non-specialists, who are unfamiliar with the concise specialized terminology; therefore, in some sections of the EHR, medical observations tend to be circumscribed.

Finally, the distribution of the additional features shows that only a minor portion of the variants are *eponyms* (0.84%). *Trade names*, on the other hand, account for more than 10% of the terms. The relatively high proportion of *trade names* can be seen as an effect of the clinical specialty of the dataset, as the therapy of endocrine diseases primarily depends on medication. Arbitrary *misspellings* are only present in less than 5% of the terms; in the clinical

genre, this is a rather low figure (cf. Dalianis 2018). One reason might be that the criteria used to assess orthographic correctness were relatively lax. As long as the variation found in a term showed some kind of systematic pattern, or could be attributed to a reduction process, it was not considered a *mis-spelling*.

7.3.2 Distribution of Formal Features across the Major Semantic Groups

To evaluate whether conceptual properties pattern with particular variation processes, the five most frequent semantic groups (i.e. DISORDERS, PROCEDURES, CONCEPTS & IDEAS, CHEMICALS & DRUGS and ANATOMY) were examined more closely. For each group, the proportion of terms coded with the individual features was calculated relative to all terms belonging to this group. Table 10 provides the full results. The individual distribution of features per group is discussed in the following sections.

7.3.2.1 Disorders

Among the DISORDERS, only about one fifth of the identified terms were rated as *standard*, which is slightly below average. Almost 35% were coded as *specialized*. The proportion of terms labeled with *reduction* features is rather low: The most common type are *abbreviations*, followed by *lexical reductions*. *Grammatical reductions* are only found in 1.96% of the terms; while this is a rather low figure, it is still above average. DISORDER terms are also highly variable in the morpho-syntactical dimension: While *derivations* are rather infrequent, the proportion of *compounds* (15.26%), and especially of *paraphrases* (9.15%), is considerably above average. While the percentage of *eponyms* seems low (0.54%), it is still higher than in most other classes. *Trade names* are only present in a small fraction of the terms (0.17%), most of which relate to adverse reactions to particular drugs (e.g. *intolerantie glucophage* 'intolerance glucophage' *(Intolerance to drug (59037007))*).

Table 10: Distribution of formal features across the major semantic groups. The first column specifies the name of the semantic group. The remaining columns provide the proportion of terms coded with the individual features in percent, relative to the total number of terms belonging to this group. For comparison, the last row provides the average proportion of terms coded with this feature across the semantic groups.

	Register		Reduction			Morpho-syntax			Additional		
Semantic group	*Standard*	*Specialized*	*Abbreviation*	*Lexical reduction*	*Grammatical reduction*	*Compound*	*Derivation*	*Paraphrase*	*Eponym*	*Trade name*	*Misspelling*
DISO	20.88	34.34	7.05	5.57	1.96	15.26	2.69	9.15	0.54	0.17	2.40
PROC	11.58	32.28	12.56	13.17	4.33	16.04	0.97	5.58	0.85	0.65	1.99
CONC	42.64	25.10	6.78	2.42	0.19	5.33	6.88	7.56	0.00	0.29	2.81
CHEM	23.73	34.99	5.77	0.43	0.03	3.90	0.02	0.05	0.03	27.76	3.28
ANAT	25.32	34.25	12.93	9.31	1.14	12.23	2.27	0.05	0.27	0.00	2.22
Average	24.83	32.19	9.02	6.18	1.53	10.55	2.57	4.48	0.34	5.77	2.54

The distribution of morpho-syntactical features clearly reflects the usage of DISORDER terms in clinical documentation, which is influenced by switches between the expert and lay perspective, as well as changes in information status. As lay speakers are not familiar with the specialized terminology, they describe their symptoms in their own terms, which accounts for the relatively high proportion of *paraphrases* (e.g. *slap op de benen* 'weak on the legs' *(Muscle fatigue (80449002))*). Since no conventionalized short forms exist for such expressions, the proportion of *abbreviations* is rather low. Only when a phenomenon has been investigated by an expert, *specialized* terms are used. These terms typically appear in the form of complex noun phrases, which comprise not only the finding, but also the anatomical location, severity or etiology (e.g. *maculair oedeem* 'macular edema' *(Macular retinal edema (37231002))*; *terminale nierinsufficiëntie* 'terminal kidney insufficiency' *(End stage kidney disease (46177005))*; *diabetische retinopathie* 'diabetic retinopathy' *(Retinopathy co-occurrent and due to diabetes mellitus (4855003))*). In contrast to the vague descriptions employed by lay people, many *specialized* terms have established abbreviations (e.g. *ni* for *nierinsuf-*

ficiëntie, *DBR* for *diabetische retinopathie*). The grammatical structure of complex noun phrases tends to be compressed as well, especially to specify the location of a finding (e.g. *souffle carotiden* 'wheeze carotids' *(Arterial bruit (30846001))*), which is evident in the relatively high proportion of *grammatical reductions*.

7.3.2.2 Procedures

Among the PROCEDURES, the proportion of *specialized* variants is at an average level (32.28%). The proportion of *standard* variants, though, is lower than in any other semantic group (11.58%). The terms in this category are extremely prone to reduction processes of all kind: With 13.17%, the *lexical reductions* are most frequent, followed by *abbreviations* with 12.56%. Most striking, though, is the high proportion of *grammatical reductions*: With 4.33%, this value is more than twice as high as the average. In the morpho-syntactical dimension, the PROCEDURES are less variable: The most frequently assigned feature is *compound* (16.04%), followed by *paraphrase* (5.58%). *Derivations*, on the other hand, are quite rare (0.97%). Finally, with 0.85%, the PROCEDURES show the highest proportion of *eponyms* among the top semantic groups.

The distribution of formal features among the PROCEDURE terms indicates that both conceptual properties and the habitual context of usage influence variation patterns. PROCEDURES are mostly referred to in those parts of the EHR that are intended for internal documentation (e.g. the *Examination* section), where the use of non-standard and reduced forms is acceptable. Moreover, PROCEDURES follow a fixed protocol and are always conducted in the same manner. This enables the use of extremely condensed forms while maintaining comprehensibility. The PROCEDURE terms also show characteristic reduction patterns: In their standard form, many complex terms combine a general noun expressing the act of examination with its object (e.g. *nierfunctieonderzoek* 'kidney function examination' *(Renal function study (44277000))*, *schildklierfunctieonderzoek* 'thyroid function examination' *(Thyroid panel (35650009))*). In the reduced form, the general noun is omitted, retaining only the object or site (e.g. *nf (nierfunctie)* 'kidney function', *sk (schildklier)* 'thyroid'). By contrast, in *radiografie van de thorax* 'radiog-

raphy of the thorax' *(Plain chest X-Ray (399208008))* and *kernspintomografie van de hersenen* 'Magnetic resonance imaging of the brain' *(Computerized axial tomography of brain (34227000))*, the headword of the term provides crucial semantic details. Therefore, the noun expressing the procedure is retained in the reduced variants. In these cases, short forms are obtained through the abbreviation of constituent words, or the omission of function words (e.g. *rx tx (x-ray thorax)*, *kst (kernspintomografie) hersenen* 'Magnetic Resonance Imaging brain').

7.3.2.3 Concepts & Ideas

Among the CONCEPTS & IDEAS, the proportion of *standard* variants is higher than in any other group (42.64%). At the same time, this group has the lowest proportion of *specialized* terms (25.10%). Reduced forms, too, are rather infrequent, with only 6.78% of *abbreviations* and 2.42% of *lexical reductions*; *grammatical reductions* were only found in 0.19% of the terms. While only 5.33% of the variants are marked as *compounds*, the proportion of *derivations* is very high (6.88%). With 7.56%, *paraphrases* are also more frequent than in most of the other semantic groups. Also, compared to the other semantic groups, the proportion of *misspellings* is substantial (2.81%).

The distinctive distribution of register and reduction features can be attributed to the relative simplicity of the terms, both at the conceptual and lexical level. Many concepts in this group belong to the general domain. The associated terms are used by laypeople and medical experts alike, both in the clinical setting and beyond. Even in their full form, these terms are rather short and have a simple orthography (e.g. *mild (Mild (255604002))*, *vaak* 'often' *(Frequent (70232002))*). Hence, there is no practical need to shorten these terms. The few reduced variants have been imported from the general domain, where they are strongly entrenched by daily usage (e.g. *d/n* 'day/night' *(Night and day (224943009))*; *gem (gemiddeld) (Averaged (371921001))*). The distribution of morpho-syntactical features clearly shows that the terms in this group mostly serve to qualify other concepts. In particular, we find a very high proportion of *derivations* (e.g. *gekleurd* 'colored' *(Color change (263715003))*, derived from *kleur* 'color'), which can act as modifiers in complex noun phrases.

7.3.2.4 Chemicals & Drugs

Among the CHEMICALS & DRUGS, the register features show an average distribution, with 23.73% of the variants coded as *standard*, and 34.99% as *specialized*. Compared to the other groups, though, the proportion of reduced variants is extremely low, with just 5.77% of *abbreviations*, 0.43% of *lexical reductions* and 0.03% of *grammatical reductions*. At the morpho-syntactical level, the terms in this group show even less variation: While the *compound* feature was still assigned to 3.90% of the variants, *derivations* and *paraphrases* are extremely rare (0.02% and 0.05% respectively). *Eponyms*, too, are very infrequent (0.03%). More than a quarter of the variants are *trade names* (27.76%), which is the highest value among the groups. Likewise, *misspellings* are more frequent than in any other group (3.28%).

The group of CHEMICALS & DRUGS clearly stands out among the major semantic groups. The conceptual scope is very confined, with the vast majority of terms referring to manufactured entities. The associated terms are strongly dominated by nominal forms, while morpho-syntactical processes are unproductive. The working mechanisms of medical drugs are rather abstract and difficult to conceptualize; for a lay person, only the effect is perceivable. Therefore, the naming practices for higher-level drug categories are typically based on the target of treatment, such as a clinical condition or physiological function (e.g. *antidepressivum* 'antidepressive' *(Medicinal product acting as antidepressant agent (36236003))*; *bloeddrukpillen* 'blood pressure pills' *(Hypotensive agent (1182007))*). For insulin in particular, category names are also based on the duration of the effect (e.g. *langwerkende insulin (Long-acting insulin (25305005))*). Among the common classes of medication, we also find established abbreviations (e.g. *ace-i (angiotensin-converting enzyme inhibitor) (Product containing angiotensin-converting enzyme inhibitor (41549009))*). Terms for more fine-grained concepts, though, are typically based on either the scientific name of the active ingredient, or the trade name of the commercial product (e.g. *simvastatine* and *Zocor (Simvastatin only product (777537002))*). The propensity to reduction processes differs between the two: Substance names are morphologically and orthographically complex; therefore, *abbreviations* are common, especially for substances prescribed against common disorders of the

cardiovascular system and antidiabetica (e.g. *simva* for *simvastatine*; *mt* for *metformine (Product containing metformin (109081006))*). Among commercial names, the potential for variation seems lower in general. One possible explanation is that, for reasons of branding, *trade names* are designed to be easy to memorize and use, even among laypeople; therefore, there is no need to coin reduced forms. However, with regard to the reduction potential, there is a clear effect of domain pertinence: The *trade name* terms referring to drugs that are not related to diabetes are typically kept in their full form. For antidiabetica, though, in particular insulin products, abbreviations of trade names are quite common (e.g. *lev* for *Levemir (Product containing only insulin detemir (776342000))*; *IT* for *Insulatard (Product containing only isophane insulin (776415000))*; *hum reg* for *Humuline Regular (Product containing short-acting insulin (325013000))*). In general, the proportion of *misspellings* is quite high. Possibly, this is an effect of the arbitrariness of commercial names: In contrast to terms from the scientific nomenclature, the spelling cannot be inferred by domain knowledge, but must be memorized. As a result, variants with equivalent phonology, but incorrect orthography are frequent among product names (e.g. *Kayexalat* and *Kayaxalate* instead of *Kayexalate (Product containing calcium polystyrene sulfonate (346361003))*). Interestingly, some spelling mistakes are systematic, in that they mirror the orthographical alternations in neoclassical terminology (e.g. *Glucofage* instead of *Glucophage (Product containing metformin hydrochloride (325271001))*).

7.3.2.5 Anatomy

About one quarter of the variants referring to body parts were judged as *standard*, and 34.25% as *specialized*. The terms from this group are very prone to reduction processes. In particular, with 12.93%, the ANATOMY terms show the highest proportion of *abbreviations* among the major semantic groups. Compared to the other groups, *lexical reductions* are very frequent as well (9.31%), whereas *grammatical reductions* are relatively rare (1.14%). The distribution of morpho-syntactical features shows a clear dominance of nominal forms: *Compounds* are most strongly represented with 12.23%; about 2.27% of the variants are *derivations*, which is a little below average. By contrast, with just 0.05%, the proportion of *paraphrases* is lower than in

any other group. Only a small fraction of the terms are labeled as *eponyms* (0.27%). Unsurprisingly, *trade names* do not occur at all in this group. Finally, with 2.22%, the proportion of *misspellings* is slightly below the average value.

Overall, the proportion of register-related features shows about average values. At a closer look though, the degree of terminological specialization seems to be co-determined by conceptual properties: In particular, perceptual accessibility influences the prevalence of *specialized* variants: For tangible body parts, which are visible to the outside, most terms are lay variants (e.g. *buik* 'belly' *(Abdomen structure (113345001))*). Similarly, for major inner organs, whose function is common knowledge, non-specialized forms prevail (*long* 'lung' *(Lung structure (39607008))*). By contrast, for concepts relating to body parts whose knowledge requires a scientific background, such as blood vessels, mostly specialized variants were identified (e.g. *arteria iliaca communis (Common iliac artery structure (73634005))*).

The majority of concepts in this group are concrete entities, which affects the morpho-syntactical variability of the terms. Most of the terms are noun phrases; compounding is particularly frequent to form variants that provide further detail for a base concept (e.g. *pancreaskop* 'pancreas head' *(Pancreas part (119218006))*; *rechterthoraxhelft* 'right half of thorax' *(Entire right thorax (362682009))*). Besides, derivation is a productive mechanism to form adjectives that serve as modifiers for terms from other categories; however, this process is limited to terms based on neoclassical roots (e.g. *abdominaal (Abdominal (277112006))*; *pulmonair (Pulmonary (264164005))*). By contrast, verbal derivations are not found at all in this group. Many terms relate to body parts that are investigated as part of a routine check-up. As their examination follows a fixed protocol, the proportion of *abbreviations* is high, also among derived terms (*abd* for *abdominaal*, *pulm* for *pulmonary*). Likewise, *lexical reductions* are relatively frequent. In particular, among complex noun phrases relating to fine-grained concepts, the grammatical head of the phrase, which expresses the general anatomical category, tends to be omitted (e.g. *carotis interna* instead of *arteria carotis interna*). *Grammatical reductions*, on the other hand, are rare; among the few cases are reduced preposi-

tional expressions, where one body site is specified by another (e.g. *lateraaltak cx* 'lateral branch circumflex' *(Structure of left posterior lateral branch of circumflex branch of left coronary artery (57823005))*).

7.4 Conclusion

The formal annotation of the terms identified in the dataset reveals the prevalence of typical sublanguage properties associated with the clinical domain. The terminology is dominated by complex nominal phrases; among these, we find a high proportion of non-native lexemes, especially terms derived from neoclassical roots, and reduced forms. The high proportion of terms rated as non-standard illustrates the discrepancy between term representation in structured knowledge sources, and their usage in clinical practice.

Between the semantic groups, there are distinctive differences in the feature distribution. These differences provide evidence that *conceptual properties*, as well as the *semantic constellations* in which a concept typically occurs, influence the prevalence of formal features. In general, conceptual accessibility patterns with vernacular variants. By contrast, for concepts whose comprehension requires specialized knowledge, the proportion of specialized variants tends to be higher. Moreover, the concreteness influences the morpho-syntactical variability: For concepts that have a prototypical manifestation, such as body parts, noun forms prevail; for phenomena, on the other hand, which might be idiopathic in nature and whose verbalization depends on subjective experience, descriptive paraphrases are more common. Besides, there is an effect of the combinatorial potential: Terms from the groups of ANATOMY, as well as CONCEPTS & IDEAS, are frequently used to qualify other concepts, e.g. with regard to their location or severity. Therefore, the proportion of derived forms is higher. Moreover, the *habitual context of usage* determines which types of variants are acceptable. For instance, PROCEDURES are very prone to reduction processes. One reason is that they are typically referred to in informal parts of the EHR, where the use of non-standard forms is more common. Besides, many PROCEDURES are complex

concepts, whose constituents always occur in fixed constellations. The fixedness of the combination legitimates the use of simplified constructions: Either the relation between the PROCEDURE and its object is left underspecified, or the term denoting the act of the PROCEDURE itself is omitted completely.

From a methodological perspective, the annotation with formal features could certainly be improved. In the absence of shared community standards, the design of a feature set to operationalize variation types was challenging. While terminological theory has brought forth typological classifications of variation processes, these are not concrete enough to be directly translated into a tag set for the clinical sublanguage, especially since the prevalent term types depend strongly on the domain. The annotation aimed to cover a broad range of features and variation processes at different levels. To keep the workload feasible, the number of features had to be limited, partly at the expense of more fine-grained distinctions, which could have produced additional insights. For example, judging from the results, a more detailed distinction between different types of nominals, especially that of premodified noun phrases and prepositional phrases, might have led to the discovery of additional variation patterns.[10] Another methodological issue was the definition of criteria for feature assignment. For instance, the encoding of a term's specialization was based on the distinction between native and non-native lexical elements. Of course, this is only a crude approximation. In fact, some neoclassical terms (e.g. *diabetes*) may be widely known to the general public; on the other hand, some complex terms consisting entirely of native roots may be more difficult to conceptualize (e.g. *wervelslagpijn* 'spinal pain on percussion'). For a representative assessment, terminological specialization should be rated in an experimental setup, involving both medical experts and lay people. Of course, this would lead to an entirely new annotation project.

[10] For a closer examination, a small set of noun phrases relating to DISORDERS and PROCEDURES was annotated with PoS tags in a separate study. The distribution of PoS sequences across the semantic groups suggests that, in complex noun phrases, the conceptual properties of the constituents, as well as their semantic relations, might influence preferences for particular phrase types. The full results are presented in Grön, Bertels, and Heylen (2018a).

Overall, though, the formal annotation produced valuable insights, which enabled the quantification of variation types in the data. The results illustrate the diversity of variation processes at different levels, which interact with each other, as well as with conceptual properties and contextual factors. Given the complexity of these interactions, though, the validation of variation patterns by statistical means is problematic for such a large and heterogeneous term sample. Therefore, the final part of this thesis will focus on a smaller set of representative concepts for the statistical modeling of term variation.

Part III

Chapter 8: Modeling Clinical Term Variation

The final goal of this thesis is to model term variation by statistical means. By leveraging the findings from the previous chapters, a classification experiment, consisting of four different tasks, was conducted. These tasks served to validate variation patterns in various configurations, and to compare the impact of conceptual and contextual influences on term selection.

While the first section presents the composition of the sample of concepts and terms included in the experiment (Section 8.1), the second section describes the set of conceptual properties and context factors that were used as predictors in the classification tasks (Section 8.2). The next section formulates the general research questions underlying the experiment, and gives an overview of the parameter constellations in the individual tasks (Section 8.3). Following an outline of the methodological setup (Section 8.4), the results of the experiments are presented (Section 8.5). The final section summarizes the findings and discuss potential applications in terminology management and clinical NLP (Section 8.6).

8.1 Composition of the Concept and Term Sample

As the results of the annotation at the concept level showed, the ratio between concepts and associated term variants has a highly skewed distribution; a small number of concepts accounts for a large part of the variants encountered in the dataset (cf. Section 6.3.2). On the other hand, the analysis of the formal term features illustrated the intricacies of different variation processes operating at separate levels, and sometimes interacting with each other (cf. Section 7.3). If the entire sample of concepts and terms would be included in a statistical model, individual effects and their interactions would not be

discernible; therefore, the classification experiment focused on a small, but representative sample.

The final sample consisted of 25 concepts, including five concepts for each of the five major semantic groups (i.e. DISORDERS, PROCEDURES, CONCEPTS & IDEAS, CHEMICALS & DRUGS and ANATOMY). For each semantic group, the concept sample was composed in the following manner: First, the concepts within each semantic group were sorted by their absolute frequency (i.e. the number of occurrences in the annotated part of the dataset) and by the number of variants that had been linked to the concepts. The 30 highest-ranking concepts were manually inspected. The final selection was made according to the following principles: Firstly, the selected concepts should cover the entire semantic spectrum of the group: For example, for ANATOMY, the final sample included one major morphologic region (*thorax*), one limb (*leg*), one inner organ (*thyroid*), one blood vessel (*left internal carotid artery*) and one body product (*urine*). Secondly, the concepts were chosen such that, if possible, the full range of variation processes could be evaluated. To ensure this would be the case, the number of associated variants, and the number of feature alternations observed among them, should be as high as possible. For example, the concept *Renal function study (44277000)* has the sixth-highest frequency among the PROCEDURES (2,089 occurrences). The associated terms show variation in different dimensions, including the conceptual (*controle van de nierfunctie* 'control of the kidney function' vs. *nierfunctie* 'kidney function'), the denominative (*nierfunctie* vs. *nf*) and the linguistic level (*nier functie* vs. *nierfunctie*). This makes it a good candidate to investigate the effect of different factors on individual variation processes, so that it was included in the final concept sample. The concept *Diabetic diet (160670007)* is on rank ten among the most frequent PROCEDURES (1,570). However, it only has three associated variants, which differ merely in orthography (*diabetesdieet, diabetes dieet, diabetes-diet*). Thus, this concept it less suited for the investigation of different variation processes; therefore, it was discarded in favor of a less frequent candidate, which displays more variation.

Table 11 – Table 15 list the final concept samples for each semantic group. On average, each concept has 5.12 associated variants; the mean number of

feature alternations among these variants is 2.84. All occurrences of these concepts were retrieved from the annotated part of the dataset. Together, they form a sample of 28,520 mentions, which served as training and testing data for the classification experiment.

Table 11: Sample of concepts and terms for the group DISORDERS, sorted by concept identifier. The first and second column list the unique identifier and the PT of the concept in SNOMED CT. The third column gives examples of the term variants associated with this concept. The final two columns specify the number of term variants, and the number of feature alternations observed among these variants.

SCTID	PT	Examples of the associated variants	Number of variants	Number of feature alternations
4855003	*Diabetic retinopathy*	*diabetische retinopathie* 'diabetic retinopathy' *diabetesretinopathie* 'diabetes retinopathy' *DRP (diabetische retinopathie)* 'diabetic retinopathy'	5	3
42399005	*Renal failure syndrome*	*nierinsufficiëntie* 'kidney insufficiency' *nierfunctieachteruitgang* 'kidney function decrease' *ni (nierinsufficiëntie)* 'kidney insufficiency'	4	3
44054006	*Diabetes mellitus type 2*	*type-2-diabetes* *diabetes type 2* *DM (diabetes mellitus) II*	6	3
45007003	*Hypotension*	*hypotens* 'hypotensive' *lage tensies* 'low tensions' *lage BD (bloeddruk)* 'low blood pressure'	4	4
312975006	*Microalbuminuria*	*microalb (microalbuminurie)* 'microalbuminuria' *malbie (microalbuminurie)* 'microalbuminuria'	2	2

Table 12: Sample of concepts and terms for the group PROCEDURES, sorted by concept identifier. The first and second column list the unique identifier and the PT of the concept in SNOMED CT. The third column gives examples of the term variants associated with this concept. The final two columns specify the number of term variants, and the number of feature alternations observed among these variants.

SCTID	PT	Examples of the associated variants	Number of variants	Number of feature alternations
26046004	*Cardiovascular stress test using bicycle ergometer*	*cycloergo (cycloergometrie)* 'cycle ergometry' *cyclo ECG (electrocardiografie)* 'cycle electrocardiography' *fietsproef* 'bike test'	7	3
35650009	*Thyroid panel*	*schildkliertesten* 'thyroid tests' *skf (schildklierfunctie) onderzoek* 'thyroid function examination' *skfie (schildklierfunctie)* 'thyroid function'	4	3
44277000	*Renal function study*	*controle van de nierfunctie* 'control of the kidney function' *nierfunctie* 'kidney function' *nf (nierfunctie)* 'kidney function'	3	3
59108006	*Injection*	*inj (injectie)* 'injection' *inspuiting* 'injection' *inspuiten* 'inject'	6	4
276342005	*Ophthalmological and optical investigations*	*consult oftalmo (consultatie oftalmologie)* 'consultation ophthalmology' *ophthalmologische controle* 'ophthalmological controle' *oogfctieonderzoek (oogfunctieonderzoek)* 'eye function examination'	9	6

Table 13: Sample of concepts and terms for the group CONCEPTS & IDEAS, sorted by concept identifier. The first and second column list the unique identifier and the PT of the concept in SNOMED CT. The third column gives examples of the term variants associated with this concept. The final two columns specify the number of term variants, and the number of feature alternations observed among these variants.

SCTID	PT	Examples of the associated variants	Number of variants	Number of feature alternations
2603003	*Secondary*	*secundair* 'secondary' *sec (secundair)* 'secondary' *verwikkeld met* 'complicated by'	3	3
62459000	*Chronic persistent*	*persisteren* 'persist' *persisterend* 'persisting' *aanslepend* 'protracted'	4	3
73775008	*Morning*	*ochtendlijk* 'morning' (adj.) *vm (voormiddag)* 'morning' *matinaal* 'matinal'	9	4
255604002	*Mild*	*matig* 'moderate' *discreet* 'discrete' *beperkt* 'limited'	6	1
398232005	*Drug dose*	*medicatiedosis* 'medication dose' *dosis* 'dose' *dos (dosis)* 'dosis'	6	2

Table 14: Sample of concepts and terms for the group CHEMICALS & DRUGS, sorted by concept identifier. The first and second column list the unique identifier and the PT of the concept in SNOMED CT. The third column gives examples of the term variants associated with this concept. The final two columns specify the number of term variants, and the number of feature alternations observed among these variants.

SCTID	PT	Examples of the associated variants	Number of variants	Number of feature alternations
7947003	*Product containing aspirin*	*asp (aspirine)* 'aspirin' *asp jr (aspirine junior)* 'aspirin junior' *cardioaspirine* 'cardio aspirin'	16	2
108548008	*Product containing bisoprolol fumarate*	*Co-Bisoprolol* *co-bis (Co-Bisoprolol)* *Emcoretic*	6	1
108575001	*Product containing lisinopril*	*Lisinopril* *Lisinopril Sandoz* *Zestril*	3	1
125703000	*Human insulin analog product*	*insuline analogen* 'insulin analogue' *analooginsuline* 'analogue insulin' *analogen* 'analogue'	3	2
320031002	*Product containing precisely atorvastatin 40 milligram/1 each conventional release oral tablet*	*atorvastatine* 'atorvastatin' *atorva (atorvastatine)* 'atorvastatin' *Lipitor*	4	1

Table 15: Sample of concepts and terms for the group ANATOMY, sorted by concept identifier. The first and second column list the unique identifier and the PT of the concept in SNOMED CT. The third column gives examples of the term variants associated with this concept. The final two columns specify the number of term variants, and the number of feature alternations observed among these variants.

SCTID	PT	Examples of the associated variants	Number of variants	Number of feature alternations
51185008	*Thoracic structure*	*thorax* *tx (thorax)* *borst* 'chest'	5	5
58379002	*Left internal carotid artery*	*carotis interna links* 'carotis interna left' *art (arteria) carotis interna links* 'arteria carotis interna left' *ACI li (arteria carotis interna links)* 'arteria carotis interna left'	3	2
61685007	*Lower limb structure*	*been* 'leg' *onderste ledematen* 'lower limbs' *OL (onderste ledematen)* 'lower limbs'	3	2
69748006	*Thyroid structure*	*schildklier* 'thyroid' *sk (schildklier)* 'thyroid' *thyroïd* 'thyroid'	3	4
122575003	*Urine specimen*	*urinestaal* 'urine specimen' *ur (urine) staal* 'urine specimen' *urines*	4	4

8.2 Operationalization of the Predictors

According to theories of terminology, term choices can be motivated by conceptual properties, as well as cognitive and contextual factors, whereby the precise nature of these factors, as well as their relative importance depends strongly on the domain (cf. Section 2.2). However, not all of these factors could be investigated with the dataset at hand. The following sections give an overview of the potential factors that may influence variation in clinical term usage, explain which factors were included in the classification experiment and how they were operationalized.

8.2.1 Conceptual Properties

Conceptual properties could affect term choices in various ways. As the output of the annotation studies indicates, it is likely that the semantic nature co-determines variation patterns. For example, concepts that tend to be used as modifiers (e.g. CONCEPTS & IDEAS), have a higher potential to show variation at the morpho-syntactical level. To validate this effect, a semantic variable was included in the statistical model. The value of this variable was based on the UMLS semantic groups, which had already been assigned in the concept annotation stage (cf. Section 6.3.1.2.2).

Moreover, conceptual complexity could influence term variation, especially with regard to register. For example, it is conceivable that, among the concepts in the group of anatomy, those concepts that are perceptually salient (e.g. *Lower limb structure*) pattern with lay terms, whereas those that can only be visualized and examined by specialized means (e.g. *Left internal carotid artery*), tend to be expressed in specialized terms. Therefore, it was considered to implement a measure of conceptual complexity based on the position of the concept in the UMLS hierarchy. The UMLS is organized in a tree structure; the position of an individual concept in the semantic tree can be retrieved from the UMLS browser (National Library of Medicine 2019b). The tree number reflects the depth at which a concept is situated. For example, *Lower limb structure* is located in the sub-tree A2.1.5.2, i.e. three steps down from the group node. However, at a closer look, it became evident that,

due to the relative coarseness of the semantic classification, the tree system does not allow for much differentiation. For example, in the group relating to ANATOMY, four out of the five included concepts are located at the same level (i.e. 3 levels below the top node); only the concept *Urine specimen* is situated at a higher level (2 levels below the top). Among the CONCEPTS & IDEAS, all concepts occupy the same level, namely 3 steps under the top node. Evidently, a measure based on the UMLS tree depth would provide little additional insight. Another option would have been to use the level in the original SNOMED CT hierarchy. However, while this would have enabled more fine-grained distinctions, it also would have caused incompatibilities with the semantic group feature. For example, *Drug dose*, which belongs to CONCEPTS & IDEAS in the UMLS, is classified as a *Qualifier*, as well as a *Finding* in SNOMED CT. Eventually, no predictor representing conceptual complexity was included.

8.2.2 Cognitive Factors

Cognitive effects on term selection can manifest themselves at the collective as well as the individual level. At the collective level, term preferences may depend on the local dialect or institutional conventions. However, the data under analysis was provided by a single hospital; for a cross-institutional comparison, a comparable corpus from another hospital would have been required. The only other Dutch clinical corpus available at the time of this research was the corpus compiled at the Erasmus University Medical Center (EMC; cf. Afzal *et al.*, (2014)). In a previous study, this corpus had been combined with the dataset under analysis for the evaluation of a more general NLP task, namely PoS tagging (Grön, Bertels, and Heylen 2018a). However, the EMC corpus is only annotated with regard to contextual properties, but not at the concept level. Moreover, it is a very heterogeneous collection, comprising, among others, radiology reports and notes written by general practitioners; this calls the comparability into question. Hence, while the comparison would have been interesting from a dialectological point of view – one dataset being composed in the Flemish, and the other in the Netherlandic variety of Dutch – it was not deemed feasible.

Cognitive influences can also emerge in the term preferences of individual practitioners. For example, preferences for a particular form can be shaped by the country of origin, or the academic institution where a clinician received their medical education. However, following the confidentiality agreement with the UZ Leuven, the only available information on the authors of the EHRs were numerical IDs. For a meaningful interpretation of sociolinguistic factors, though, background information on the authors would have been required. Another problematic aspect for the modeling of individual preferences is the high proportion of textual overlap between the EHRs. At the beginning of a consultation, clinicians routinely pull up the report from the previous encounter to put the patient's case into context. If no update is required, the corresponding text portions are simply copied and pasted into the EHR of the current consultation; obviously, this would blur individual preferences. For methodological reasons, the investigation of the effect of personal preferences on term choices was thus not possible.

To quantify the degree of textual overlap between consecutive EHRs relating to one patient, the average number of new token sequences per section was calculated. For each case history (i.e. the complete set of EHRs associated with one patient), all text snippets from one section were extracted from the individual EHRs. Then, the relative overlap was determined by comparing pairs of text snippets in chronological order. Starting with the first and the second snippet, the longest common substring (LCS) between the two was identified.[11] Then, the length of the LCS was subtracted from the length of the second snippet. Finally, the relative gain in new tokens was calculated by dividing the length difference between the two snippets by the total length of the second snippet. The relative gain thus gives an indication of the proportion of new tokens in a snippet.

Table 16 shows the average proportion of new tokens across the main sections. Evidently, the average values vary strongly between the sections, reflecting differences in the communicative function of the sublanguage. As

[11] To compute the LCS, a Python implementation of the Suffix Tree Algorithm was used (https://pypi.org/project/suffix-trees/).

expected, the *History* is a very conservative section, where, on average, almost 90% of the tokens are copied over from earlier EHRs. By contrast, the *Comments* are much more dynamic, with almost 90% new tokens. This implies that a substantial portion of the words in an EHR could have been written by another person than that whose ID is attached to the file. While these are noteworthy findings in themselves, they clearly show that, even if patterns of individual preferences would emerge in the dataset, they would be a mere chance effect. Therefore, no predictors reflecting cognitive factors were included in the experiment.

Table 16: Average gain of new tokens across sections. The first column specifies the section name. The second column provides the mean proportion of new tokens that are added to a section per consultation, relative to the absolute length of the section.

Section	Average proportion of new tokens added per EHR in %
Anamnesis	23.60
Comments	88.57
Complaints	80.76
Conclusion	83.59
Diet	9.11
Examination	45.75
Eye Report	66.24
History	10.36
Medication	26.15
Therapy	72.99

8.2.3 Contextual Factors

Finally, in specialized communication, term choices can be influenced by a range of context factors. The question of what constitutes context in general, and which types of context are relevant for terminology usage, has been subject to some debate: Depending on the definition, context may involve the immediate co-text (i.e. adjacent words in a piece of writing), the setting and participants of a communicative situation, as well as historical background and cultural conventions (Faber and León-Araúz 2016).

Faber and León-Araúz (2016) propose a typology of context based on the distinction of scope (local and global) and type (syntactic, pragmatic and

semantic). Local context is understood at the textual level (i.e. neighbor words and their properties), while global context comprises extra-textual factors (e.g. the formality of the speech situation). Furthermore, at each level, they distinguish between syntactic, pragmatic and semantic types of context (e.g. grammatical dependencies, register, predicate-argument structure). However, given the close interaction of different types of context, it might not always be possible to maintain such a clear distinction. In particular, specialized languages contain a high proportion of MWEs, i.e. fixed sequences of words or word types, which occur with above-average frequency. In such terms, habitual semantic configurations tend to pattern with particular syntactic structures.[12]

For the current experiment, two predictors were included to model the local and global context. Firstly, the local level (i.e. the micro-context) was represented by the neighbor tokens within a window of three (i.e. three words to the left, and three words to the right of a variant). The decision to use a relatively small window size was based on the findings of earlier studies, which suggest that, for most classification problems, narrow contexts are better suited when dealing with the clinical genre. For example, Tao, Filannino, and Uzuner (2017) report that, for the extraction of medication information, robust results were achieved with a context of two tokens to the left and right of the target word; increasing the window to up to five tokens on each side did not improve the performance of the classifier. One reason for this effect is the rather low level of syntactic complexity in clinical writing. As illustrated by the characterization of the different EHR sections (cf. Chapter 6), many sublanguages employ mainly fragmentary constructions or mere enumerations. As long-distance dependencies between the words are rare, the immediate context is most informative. However, using a window of only two tokens would be too restrictive in our case, since the target terms can appear as part of MWEs. Since stop words were not removed

[12] To investigate the interaction of the syntactic and semantic structure of clinical MWEs in more detail, a separate study had been carried out based on the dataset under analysis (Grön, Bertels, and Heylen 2018b).

from the data,[13] many MWEs span more than three tokens. Thus, a window of only two context tokens would miss important syntagmatic relations. For example, the term *been* 'leg' appeared as part of the prepositional phrase *lipodistrofie thv van re (rechter) been* 'lipodystrophy at the right leg'.

Secondly, the EHR sections were used to represent context at the global level. As illustrated by the detailed description presented earlier, the sublanguages used in the different sections vary with regard to their communicative function, which affects their semantic structure, as well as their degree of specialization and well-formedness (cf. Chapter 6). The section of occurrence was thus employed as a predictor representing the semantic and pragmatic macro-context.

8.3 Overview of the Classification Experiment

The classification experiment served to validate the hypotheses presented earlier (cf. Section 4.2), namely that sublanguage properties can be leveraged to predict variation processes.

The experiment consisted of four tasks: The first three tasks investigated whether context factors can be used to predict the variation encountered in the surface form. To evaluate in how far variation processes can be isolated from the individual instance (i.e. the token level), the level of abstraction was increased over the tasks. Task 1 started with the most basic relation: Here, the aim was to predict which variant would occur for a given concept based on the context features. Task 2 moves the classification problem to the type level: Rather than predicting an individual variant for a given concept, the aim was to predict the term type, i.e. the configuration of formal features present in the variant (e.g. *standard* and *specialized,* or *standard* and *abbre-*

[13] The reason why all stop words were left in place is that, as shown by the sublanguage descriptions presented earlier (cf. Chapter 5), their presence or absence is characteristic of the register of the sublanguage. The use or omission of function words, such as determiners, indicates a relatively high level of grammaticality. Stop words can thus serve as cues for the prediction of particular term types.

viation, etc.). Task 3 evaluated whether the occurrence of variation processes could be isolated completely from the underlying concept, and modeled as an effect of context alone: Here, the aim was to predict the presence of individual term features (i.e. *standard* or not, *specialized* or not) in the variants associated with one semantic group. Compared to Task 2, the crucial difference is that in Task 3, the potential influence of conceptual properties was minimized. This should clarify which features can be linked to individual concepts, and which can be associated with more general effects. To compare the influence of local and global context factors, Tasks 1 – 3 were conducted in three settings: firstly, using only the micro-context (i.e. the neighbor tokens) as predictors; secondly, using only the macro-context (i.e. the EHR sections); and thirdly, using both the micro- and the macro-context.

The final task served to assess whether the effects of sublanguage features on term choices, which were established by the first three tasks, were robust enough to be modeled from the opposite angle as well. Thus, in Task 4, the conceptual properties were the target of prediction. This task evaluated whether, given the formal features of a variant encountered in a particular sublanguage, it was possible to assign this variant to a semantic group, or even a particular concept.

The aims and variables included in the four tasks are described in detail below. In addition, Table 17 gives an overview of the parameter configuration of the different tasks.

Table 17: Overview of the aims and parameter configurations of the classification tasks. The first two columns specify the number and aim of the task. The third column states the observations that were used as input data for training and testing. The last two columns provide the target and predictor variables used in the task.

Task	Aim	Observations	Target variable	Predictor variables
1a	Given a concept, predict the term variant by micro-context	Terms associated with one concept	One of multiple term variants	Neighbor tokens
1b	Given a concept, predict the term variant by macro-context	Terms associated with one concept	One of multiple term variants	Section

1c	Given a concept, predict term variant by micro- and macro-context	Terms associated with one concept	One of multiple term variants	Neighbor tokens and section
2a	Given a concept, predict the term type by micro-context	Feature configurations of the terms associated with one concept	Multiple formal features	Neighbor tokens
2b	Given a concept, predict the term type by macro-context	Feature configurations of the terms associated with one concept	Multiple formal features	Section
2c	Given a concept, predict the term type by micro- and macro-context	Feature configurations of the terms associated with one concept	Multiple formal features	Neighbor tokens and section
3a	Given a semantic group, predict the term features by micro-context	Occurrences of individual formal features in one semantic group	One formal feature	Neighbor tokens
3b	Given a semantic group, predict the term features by macro-context	Occurrences of individual formal features in one semantic group	One formal feature	Section
3c	Given a semantic group, predict the term features by micro- and macro-context	Occurrences of individual formal features in one semantic group	One formal feature	Neighbor tokens and section
4a	Given the formal term features, predict the semantic group of the underlying concept	All formal features of the terms belonging to a semantic group	One of multiple semantic groups	Multiple formal features
4b	Given the formal term features, predict the underlying concept	All formal features of the terms associated with one concept	One of multiple concepts	Multiple formal features

8.3.1 Task 1: Prediction of Term Variants by Context

This task aims to predict the occurrence of a particular variant for a given concept in context. The underlying assumption is that syntagmatic relations within a text (i.e. the micro-context), as well as pragmatic circumstances (i.e. the macro-context) can influence term selection.

The input data used in this task was the list of annotated entities that had been linked to the 25 included concepts; the entities were associated with the individual concept and aggregated by the five semantic groups (DISORDERS, PROCEDURES, CONCEPTS & IDEAS, CHEMICALS & CRUGS and ANATOMY). Three different settings were used to evaluate the informativity of the different predictor types: First, only features based on the micro-context (i.e. the neighbor tokens) were used as predictors (Task 1a); next, only the macro-context, (i.e. the EHR section in which an entity occurred), was used (Task 1b); finally, both the micro- and macro-context were combined (Task 1c).

The main hypothesis was that the informativity of the two predictor types would vary across the semantic groups. In particular, the predictive power should depend on the types of variation that are typically observed among the concepts associated with a group, and the dispersion of the concepts across different sections: For instance, for CONCEPTS & IDEAS, which show a high degree of variation at the morpho-syntactical level (cf. Section 7.3.2.3), the micro-context should be the stronger predictor. Likewise, for CHEMICALS & DRUGS, whose occurrences are concentrated in few sections (cf. Section 6.3.1.2.2), it is unlikely that the macro-context will have much predictive value; presumably, the immediate lexical context will be more informative. On the other hand, the macro-context should be more reliable for the classification of those concepts that have a high potential for register-related variation and reduction processes, such as DISORDERS. Besides, the semantic dependency of a concept could influence the significance of the predictor types: For concepts that habitually occur as part of a fixed semantic constellation, and hence tend to appear in conventionalized MWEs, the neighbor words should be most distinctive. For example, such an effect was expected for concepts from the group of ANATOMY, which frequently occur either as the object of a PROCEDURE, or the location or a DISORDER (cf. Sections 7.3.2.1, 7.3.2.2 and 7.3.2.5).

8.3.2 Task 2: Prediction of Term Types by Context

The second task followed up on the context-based modelling, but moved beyond the lexical level. Rather than predicting the occurrence of a single variant, the aim was to predict the occurrence of a term type, i.e. the configuration of formal features present in the variant.

As in Task 1, the input data consisted of the list of term observations, labeled with the concept ID from SNOMED CT and sorted by the five semantic groups. However, rather than by the individual variant, the instances on the list were represented by the set of features characterizing this variant.

In this and the following tasks, only the features reflecting register switches (*standard* and *specialized*), reductions (*abbreviation* and *lexical reduction*) and morpho-syntactical alternations (*derivation, compound* and *paraphrase*) were included. The remaining features (*grammatical reduction*, *eponym, trade name* and *misspelling*) were not used as variables, due to the following reasons: With only five occurrences, *grammatical reductions* were far too infrequent to be included. While *eponyms* do not occur at all among the included terms, *trade names* only occur within one semantic group (CHEMICALS & DRUGS), such that, among the included terms, this feature would actually represent a semantic group, rather than a formal property. *Misspellings*, on the other hand, are too arbitrary; modeling them as a systematic alternation would not be insightful in the current study.

Like Task 1, Task 2 was evaluated in three different settings to compare the predictive power of the context factors: firstly, using only the micro-context (Task 2a); secondly, using only the macro-context (Task 2b); thirdly, using both micro- and macro-context in a combined model (Task 2c).

Building up on the premises of Task 1, the hypothesis was that the strength of the predictors would vary. In particular, their informativity should depend on the characteristic variation types found in a semantic group. Presumably, the micro-context would be most informative for the classification of concepts that tend to show variation at the morpho-syntactical level, or which tend to

occur as part of fixed MWEs. Conversely, the macro-context should be a strong predictor for concepts whose associated variants differ in register. However, since in this task, the variation features were disassociated from the lexical forms, the assumption was that syntagmatic relations would be less reliable cues. Hence, the effect of the micro-context should be attenuated, while we should see an increase in the relative contribution of the macro-context.

8.3.3 Task 3: Prediction of Term Features by Context

The aim of the third task was to investigate whether variation processes could be modeled as an abstract phenomenon, i.e. irrespectively of the underlying concept. Other than in Task 2, where the observations were associated with the individual concepts, this task was set up at the group level. Given only the semantic group and the context, the presence or absence of an individual feature should be predicted. This should shed light on which features can be tied to individual concepts, and which are an effect of more general properties of the sublanguage. Moreover, the task evaluated whether the influence of the different context levels differed across the individual features (e.g. *standard*, *abbreviation* or *derivation*), and across the feature groups (i.e. register switches, reduction processes and morpho-syntactical alternations).

In contrast to the previous tasks, the input data was not based on the terms associated with one concept, but consisted of the observations of individual features within a semantic group. One observation was thus represented by the presence or absence of an individual formal feature (e.g. *standard* or not) along with the context features.

As in the previous tasks, three iterations were conducted to compare the effect of using only the micro-context (Task 3a), only the macro-context (Task 3b), or both (Task 3c).

This task further extended the hypothesis investigated in Task 1 and Task 2. The general assumption was that the macro-context should be more informa-

tive for the prediction of features reflecting register switches and reduction processes, while the micro-context should be more reliable for the classification of morpho-syntactical alternations. However, the alternations were modeled in isolation in order to analyze them at the type level, rather than predicting single instances. As no information about the concept was provided, the predictability of the features should vary, depending on whether they were caused by local constraints of the linguistic system or individual concept constellations, or by more general sublanguage properties.

8.3.4 Task 4: Prediction of Semantic Properties by Formal Features

While the first three tasks relied on context features to predict the observed form or properties thereof, Task 4 set out from the opposite angle. Based on the observation that the nature of a concept influences its propensity to show particular variation processes (cf. Section 7.4), it investigated whether, in turn, the formal features of a variant could be exploited to infer its semantic properties.

As in Task 2, the classifier was presented with a list of term observations, whereby the terms were represented by their feature constellation. In contrast to the previous tasks, though, both the identity of the concept, and the semantic group label were removed; instead, the observations were sorted by the section of occurrence (i.e. the macro-context). The task was run in two setups: In the first run, the target variable was the semantic group (Task 4a); in the second run, the target was the identity of the concept (Task 4b). The number of targets thus varied, depending on which semantic groups and concepts were present in the respective section. In both subtasks, the term type of the variant (i.e. the constellation of formal features) was used as the predictor.

The hypothesis was that the outcome would be modulated by the semantic structure of the section: In the semantically homogeneous sections (e.g. *Therapy*), which are strongly dominated by a single concept class, better results were expected than in the more heterogeneous ones. Moreover, the

general level of the scores would depend on the complexity of the classification task. In particular, a drop in performance can be expected between Task 4a, where the classifier only had to distinguish between a small number of groups, and Task 4b, where the number of possible classifications was higher.

8.4 Experimental Setup

For all classification experiments, the Random Forest Classifier (RFC; Breiman (2001)) was used in a Python implementation.[14] This classifier was chosen because earlier studies reported that it performs well in sparse-data scenarios, where the number of predictors is very high compared to the number of observations (e.g. Xu and Jelinek (2004); Matsuki, Kuperman, and Van Dyke (2016)). This is an important quality in those tasks where the neighbor tokens are used as predictors, as the number of predictors equals the vocabulary size. Moreover, RFCs are known to deal well with classification problems where collinearity between predictors is possible (Deshors and Gries 2016). This issue must be taken into account when using the formal term features as predictors, since the features could be correlated with each other.

The performance of the classifier was measured by the F1 score, i.e. the harmonic mean between precision and recall. For Tasks 1 and 2, in each iteration, one model was built per concept; then, the results were averaged across the semantic groups. For Task 3, one model was trained for each semantic group and each formal feature; for evaluation, the average values for each feature were calculated. Finally, for Task 4, one model was trained per section.

[14] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

8.5 Results

8.5.1 Task 1: Prediction of Term Variants by Context

On average, the best results in this task were achieved in the combined model, using both the local and global context. However, the results illustrate differences in the strength of the predictors across the semantic groups. In Task 1a and 1b, where the predictor types are evaluated in isolation, the micro-context was the stronger predictor for CONCEPTS & IDEAS, CHEMICALS & DRUGS and ANATOMY. However, for all three groups, the combination of both types of context (Task 1c) produced the best results in the end. By contrast, for DISORDERS and PROCEDURES, the macro-context was most informative, outperforming the combined model as well. The full results are provided in Table 18.

Overall, the hypothesis presented earlier was confirmed. The results showed a clear division between those semantic groups that vary mostly at the morpho-syntactical level, and those that are prone to show register-related types of variation. For instance, CONCEPTS & IDEAS mostly act as modifiers for other concepts (e.g. *ochtendlijke hypoglycemie* 'matinal hypoglycemia', *matige oedeemvorming* 'moderate formation of edema'); the chosen form thus mostly depends on the immediate context. Moreover, the associated variants do not vary much in register, such that the section of occurrence had little informative value. Similarly, terms relating to ANATOMY tend to be subordinate constituents in more complex terminological units, where they either serve to localize an observation (e.g. *thoracale druk* 'thoracic pressure'), or are the object of a medical action (*doppler OL (onderste ledematen)* 'Doppler ultrasound lower limbs'); the selected variant could thus best be predicted by the neighbor tokens. By contrast, for DISORDERS, the chosen variant was most accurately predicted by the section. One reason is that DISORDERS tend to be conceptually unstable, ranging from symptoms perceived by laypeople (*lage bloeddrukken* 'low blood pressures') to concise diagnoses formulated by specialists (*orthostatische hypotensie* 'orthostatic

hypotension'). The selection of a variant thus strongly depends on the point of view of the observer, and the stage in the diagnostic process. Therefore, the macro-context was the stronger predictor. Another reason might be that most of the DISORDERS in the sample are common complications of diabetes (*Diabetic retinopathy*, *Renal failure syndrome* and *Microalbuminuria*), whose presence or progress is routinely checked as part of every clinical examination. These concepts are thus both mentioned in the sections serving internal documentation (*Comments*, *Examination*), and the sections presenting diagnostic statements (*History*, *Conclusion*). Thus, as term choices depend on the formality of the section, the macro-context was more informative for this semantic group. For the PROCEDURES, an even larger gap in performance emerged, with the macro-context scoring almost 10% higher than the micro-context. However, in this group, this effect can likely be attributed to a different cause: The PROCEDURES relate to standardized actions conducted in the clinical setting, which always follow a fixed protocol. Therefore, in those sections where the use of non-standard variants is acceptable, this group has a high potential for reduction processes; the macro-context was thus a stronger predictor.

Table 18: Results for the prediction of term variants by context, averaged across the semantic groups. The first column specifies the semantic group. The remaining columns provide the F1-score achieved by using only the micro-context, only the macro-context or the combined model. The last row provides the average values achieved in each setting. In each row, the highest value is set in bold.

Semantic group	F1 micro-context	F1 macro-context	F1 combined model
DISO	0.7080	**0.7472**	0.7451
PROC	0.6697	**0.7727**	0.7207
CONC	0.5441	0.5205	**0.5443**
CHEM	0.6647	0.6583	**0.6922**
ANAT	0.7610	0.6591	**0.7685**
Average	0.6695	0.6716	**0.6942**

For CHEMICALS & DRUGS, the results are difficult to interpret. The initial hypothesis had been that, for concepts that tend to be concentrated in few sections, which is the case for this group, the macro-context would provide

little insight. However, the results show that micro- and macro-context performed almost equally in isolation, and best if both are combined.

8.5.2 Task 2: Prediction of Term Types by Context

Overall, for the prediction of the term types of individual concepts, the scores were higher than in Task 1. The increase can be explained by the fact that the decision faced by the classifier was less complex, as the number of term types observed per concept was lower than the number of associated term variants. Especially for CONCEPTS & IDEAS, as well as CHEMICALS & DRUGS, the F1 increased dramatically compared to Task 1. While the concepts in these groups have a relatively high number of associated variants (i.e. the target variable in Task 1), the number of variation types observed among them (i.e. the target variable here), is rather low. The full results are shown in Table 19.

As expected, since the variation processes were dissociated from the lexical level, the informativity of the surrounding tokens decreased overall. Instead, in Task 2b, the EHR section emerged as the strongest predictor for most groups. While there was little difference in the average performance of the three models, the scores achieved by the different predictor types still showed some modulation across the semantic groups: For all groups except ANATOMY, the macro-context produced the best results. While the combination of both context levels was overall beneficial in Task 1, this was not the case here. For two groups, i.e. for CHEMICALS & DRUGS and especially for DISORDERS, the combined model was even the weakest of all models. As an exception, for ANATOMY, the neighbor tokens remained the strongest predictors. As in Task 1, the performance gap between micro- and macro-context remained large in this group, which provides further evidence for the strong influence of local syntagmatic constraints on anatomical term selection.

Table 19: Results for the prediction of term types by context, averaged across the semantic groups. The first column specifies the semantic group. The remaining columns provide the F1-score achieved by using only the micro-context, only the macro-context or the combined model. The last row provides the average values achieved in each setting. In each row, the highest value is set in bold.

Semantic group	F1 micro-context	F1 macro-context	F1 combined model
DISO	0.7993	**0.8111**	0.7515
PROC	0.8233	**0.8888**	0.8585
CONC	0.7270	**0.7642**	0.7457
CHEM	0.9306	**0.9309**	0.9277
ANAT	**0.8497**	0.7510	0.8365
Average	0.8260	**0.8292**	0.8240

8.5.3 Task 3: Prediction of Term Features by Context

Overall, the results for the prediction of individual formal features among the variants in one semantic group were slightly lower than those achieved in Task 2, where the feature configuration of individual concepts had been the target of classification. On average, the macro-context produced the highest scores. At a slightly lower level, the micro-context and combined model performed roughly equal. The best results were achieved for the morpho-syntactical features (*derivation* and *paraphrase*), followed by the reduction feature *abbreviation*. Among the register-related features, the F1 was lower in general. Table 20 provides the full results.

The results provide some evidence for the patterning of the feature groups with different predictor types. In particular, the reduction features, whose presence is a corollary of informal sublanguages, were best predicted by section. The same trend showed with the register feature *specialized*, where, compared to the low value achieved by the micro-context, the use of the macro-context led to a jump in performance. However, such an effect did not emerge for the other register feature, *standardized*: While the macro-context still performed slightly better, the combined model achieved the best results. For the morpho-syntactical features, the assumption had been that the immediate neighbor tokens would be the strongest predictor. However, this was

only the case for one feature, *derivation*. For *compounds* and *paraphrases*, on the other hand, the EHR section proved to be more informative. With regard to the *paraphrases*, one possible explanation is that, in our dataset, the use of paraphrastic descriptions is most common in those sections reflecting a lay perspective; therefore, the macro-context was a strong predictor for this feature.

Table 20: Results for the prediction of individual term features. The first column specifies the respective feature. The remaining columns provide the F1-score achieved by using only the micro-context, only the macro-context or the combined model. The last row provides the average values achieved in each setting. In each row, the highest value is set in bold.

Feature	F1 micro-context	F1 macro-context	F1 combined model
Standard	0.6788	0.6814	**0.6893**
Specialized	0.6277	**0.7739**	0.6283
Abbreviation	0.8261	**0.8872**	0.8182
Lexical reduction	0.7704	**0.8145**	0.7696
Compound	0.7573	**0.7940**	0.7580
Derivation	**0.9381**	0.8377	0.9362
Paraphrase	0.8765	**0.8994**	0.8704
Average	0.7821	**0.8126**	0.7814

In general, the findings of this task provided further evidence for the strong influence of sublanguage properties on variation processes. For most features, the macro-context overrode the micro-context as a predictor. However, with regard to the hypothesis that the performance of different predictor types would pattern with the feature groups, the results were inconclusive.

8.5.4 Task 4: Prediction of Semantic Properties by Formal Features

On average, the results for the prediction of semantic groups (Task 4a) were better than those for the prediction of individual concepts (Task 4b). This trend is unsurprising, as the number of target classes was lower in the first setting (Task 4a). Notably, though, in one section, namely the *Examination*,

we see the opposite effect: Here, the prediction of concepts was slightly more accurate than the assignment to a semantic group. Overall, in Task 4a, the highest scores were achieved in the sections relating to *Diet*, *Medication* and *Therapy*, whereas the lowest values showed in the *History* and *Conclusion*. In Task 4b, the best results were obtained in the *Diet* and *Anamnesis*, closely followed by the *Examination*; again, the classifier performed worst in the *History* and *Conclusion*. Table 21 provides the full results.

Across the sections, there was a clear effect of semantic heterogeneity and terminological diversity, which is in line with the hypothesis presented earlier: In the *Diet*, *Medication* and *Therapy*, which concentrate on concepts from a narrow semantic range, the assignment of a group label was a rather trivial task; therefore, the F1-scores in Task 4a were very high. However, while in the *Diet* section, the score remained stable in Task 4b, in the *Medication* and *Therapy*, the performance dropped dramatically. The reason may be that, in the medication-centered sections, there is not much difference with regard to the term types associated with the individual concepts; hence, the classifier was unable to make meaningful associations. Conversely, in the *History* and *Conclusion*, the results were low overall. This can be explained by the combination of two sublanguage features: On the one hand, these sections are rather heterogeneous, covering concepts from the entire semantic spectrum. On the other hand, they are rather conservative with regard to term choices, using either canonical specialized forms or established abbreviations. Therefore, the formal term features were less informative. By contrast, in those sections that are characterized by non-standard term usage, the formal features were better suited for the semantic differentiation. Therefore, in the *Anamnesis*, *Eye Report*, *Comments* and *Complaints*, high scores were obtained in both tasks. Finally, the results for the *Examination* are exceptional, in that this is the only section where the score for the classification of individual concepts was higher than that obtained for the classification of semantic groups. Possibly, the reason is that this section documents routine procedures in a rather informal way. While the use of non-standard variants is common, these variants are strongly conventionalized. Therefore, the formal features can be tied to individual concepts, but do not generalize across semantic groups.

Overall, the results demonstrated that, within individual sublanguages, term types pattern with conceptual properties, which enables the semantic classification of unseen variants by their formal features alone. However, the quality of the results depends strongly on the diversity of the input data: In sections that are either confined to a narrow semantic spectrum, or constrained by formal standards, the classifier was unable to pick up on meaningful associations. On the other hand, in those sections that are semantically heterogeneous, and show variation at different linguistic levels, the performance was more robust.

Table 21: Results for the prediction of semantic properties across sections. The first column specifies the respective section. The second and third provide the average F1 scores for the prediction of semantic groups, and the prediction of individual concepts. The last row provides the average values for the task. In each row, the highest value is set in bold.

Section	F1 group prediction	F1 concept prediction
Anamnesis	**0.8670**	0.8188
Comments	**0.8020**	0.6863
Complaints	**0.7020**	0.6480
Conclusion	**0.5591**	0.4083
Diet	**1**	**1**
Examination	0.7455	**0.7995**
Eye Report	**0.8512**	0.8361
History	**0.5409**	0.3875
Medication	**0.9976**	0.4204
Therapy	**0.9620**	0.5057
Average	**0.8027**	0.6511

8.6 Conclusion

The final part of this thesis moved the analysis of term variation from a descriptive to an inferential level. The annotation studies presented in the earlier chapters had revealed systematic patterns in term usage across the different sublanguages within an EHR. Drawing on these findings, a series of experimental tasks was conducted. These tasks served to validate the patterns in an increasing level of abstraction, starting from the prediction of an indi-

vidual variant, over the formal type of this variant, to the presence of individual formal features. The results from Task 1 and 2 illustrate that, across semantic groups, the sensitivity to different types of context factors varies. As shown by Task 3, within a given sublanguage, it is even possible to link certain variation processes to different types of contextual influences without taking the underlying concept into account. However, the results also demonstrate the difficulty of disentangling the factors motivating term choices. Given the juxtaposition of linguistic processes operating at different levels, it is challenging to model individual variation types in isolation. Crucially, the adequate operationalization of both cause and effect is required. Still, as the results of Task 4 demonstrate, some variation patterns are robust enough to inform an even more advanced task: Within certain sublanguages, the strong association between formal features and conceptual properties enables the semantic classification of unseen term variants.

Overall, the experimental results are encouraging. They provide further evidence that term variation is far from arbitrary, but a systematic phenomenon, which can be represented and processed at the type level, rather than that of single instances. These findings have implications for both terminology management, as well as clinical NLP: As previous research suggests, the exhaustive listing of all possible variants is not a realistic, let alone efficient, strategy to deal with clinical term variation. Therefore, one possible approach could be to enrich conceptual entries with typical variation patterns, rather than adding more variants. Such patterns could, for example, take the form of reduction rules which are most likely to apply in particular contexts. At the same time, the insights gained from this experiment could inform more advanced tasks in clinical NLP. For example, clinical NER could exploit sublanguage-specific variation patterns for the mapping of non-standard variants to ontological concepts: Context-sensitive variation rules could guide the normalization of non-canonical forms, or support the resolution of ambiguous terms. Another potential area of application would be the automatic acquisition of terms from domain corpora, e.g. for the extension of existing terminologies to additional languages. Based on the patterning of variation types with conceptual properties, sublanguage features could inform the

semantic classification of term candidates, and thus support their integration into existing terminologies.

Chapter 9: Term Variation as a Function of Sublanguage Properties

To round off this thesis, this chapter summarizes the results and discusses their further implications. In the first section, the findings of the previous chapters are briefly recapitulated (Section 9.1). Next, the case study is evaluated by comparing the results against the goals set in advance (Section 9.2.1), and discussing its limitations (Section 9.2.2). Then, the implications for further research are outlined, both with regard to the management of terminological resources (Section 9.3.1), and clinical NLP (Section 9.3.2). The final conclusion is presented in Section 9.4.

9.1 Recapitulation

To position this research against the wider context, the thesis started by outlining the impact of term variation on the secondary use of health data: The mass adoption of the EHR over the past decades has led to fundamental changes in health documentation world-wide. Still, the majority of information is encoded in natural language, which is considered more efficient, expressive and flexible than standardized formats. Efficiency, expressivity and flexibility are, however, exactly the qualities that make the automatic processing of EHRs so difficult: They give way to the use of reduced, ambiguous and variable expressions, which might involve non-standard linguistic structures and non-canonical term variants. Terminological theory has long regarded variation as a systemic flaw, which could reduce the efficiency of specialized communication and induce misinterpretations. With the growing influence of socio-cognitive linguistics, though, and the increasing popularity of automatic methods for the analysis of large specialized corpora, variation came to be acknowledged as a functional concomitant of specialized discourse, which responds to the practitioners' need to express themselves in a nuanced way and a manner that is appropriate in the respective speech

situation. Subsequently, term variation came to be studied as a systematic phenomenon, which can be classified by its manifestation in the surface form, and can be linked to linguistic motivations. Crucially, to make sense of these motivations, the specific properties of specialized languages must be taken into account. As postulated by sublanguage theory, such languages are governed by constraints which might deviate from the rules of general language, and which are co-determined by the domain, as well as pragmatic factors.

However, while sublanguage theory has found wide application in the analysis of clinical language, it has never been integrated with the systematic study of term variation. By contrast, this thesis proposed to analyze term variation as a function of sublanguage properties. To undertake this step, the second part of the thesis presented a case study. First, a clinical dataset was annotated with concept identifiers; then, the individual terms were annotated with formal features. This allowed a detailed characterization of the sublanguages and variation processes present in the dataset.

As shown by the experiment presented in the third part, the combination of semantic and contextual features can predict the occurrence of individual variants or term types with surprising accuracy. Conversely, the occurrence of formal features in context can serve as a cue to infer the underlying semantics. The patterning of sublanguage properties with variation processes can thus be considered a robust phenomenon, which can be modeled by statistical means.

9.2 Evaluation of the Case Study

9.2.1 Goals and Outcomes

The presented case study set out from the observation that, although previous research provided detailed typologies of term variation in specialized languages, as well as sublanguage descriptions in various medical domains, the two lines of research have never been brought together in a comprehensive

analysis. In particular, existing studies failed to link term choices to general properties of the linguistic subsystem, such as the semantic structure and socio-linguistic factors.

To close this gap, a comprehensive clinical dataset was studied in detail. The sublanguages present in this dataset were first characterized based on their thematic focus, communicative function and stylistic properties. Then, the proportion of semantic types was quantified across the sublanguages. For an abstract description of the individual terms, a formal feature set was developed, reflecting both general and domain-specific variation processes. The experiments showed that, using such a feature set, systematic patterns between sublanguage properties and variation processes can be exposed.

Overall, the case study achieved the pre-defined goal: It presented empirical evidence that term variation in clinical communication is far from arbitrary, but related to the semantic and pragmatic differences between individual sublanguages. By developing a scheme for the formal description of term variants, it demonstrated that it is possible to isolate variation processes from the individual tokens, and model them at an abstract level. Thus, it presented a method to integrate distinctive terminological preferences and variation processes into sublanguage descriptions.

9.2.2 Limitations

While the insights gained from the case study have encouraging implications for further research (cf. Section 9.3), they have some limitations:

The significance of the results depends crucially on the quality of the operationalization scheme. However, as with all experiments involving the statistical modeling of language, the decomposition into abstract features is a simplification of both the contextual and cognitive factors that influence term choices, and the nature of the term that appears on the surface. For instance, in the case study, the EHR section was taken as a proxy of pragmatic influences. While this was the only feasible option in this case, this does not capture the complex reality of clinical communication and documentation.

Cognitive factors, too, could not be integrated due to methodological limitations. Also, some of the features that were included in the final model can only be seen as a rough approximation: For instance, the criterion used to rate a term's degree of specialization was whether it was based on a foreign root or not. Evidently, this categorization might be at odds with the cognitive reflexes of individual speakers. However, the development of a feature set that adequately reflects such properties, e.g. through the experimental rating of conceptual complexity, would have come at immense costs.

Moreover, the study did not evaluate the generalizability of the approach. The findings are based on data from a single institution, and only cover one clinical specialty. As explained earlier (cf. Section 8.2.2), it was not feasible to carry out a cross-institutional comparison due to the lack of a comparable dataset. Likewise, while an evaluation across clinical domains had been envisioned at the start of the project, this could not be realized due to the limited timeframe. Collecting the raw data and obtaining permission for its analysis proved more complex in practice than anticipated, such that the annotation itself only started in the second half of this dissertation project. Given the generally slow progress of the annotation, it was decided that it would be more insightful to annotate a larger number of EHRs from one specialty, which would provide a comprehensive view on term usage within this domain, rather than annotating smaller datasets from multiple specialties, but just scratching the surface of their terminological richness.

Another goal that could not be achieved as planned was the combination of the insights gained from this case study with the results obtained by our project partner, LIIR. As the research conducted by LIIR was focused on the extraction of temporal and spatial relations, the original outline of the project had foreseen to synthesize the two approaches in a final step: The insights gained about term variation should be combined with those concerning the relations between entities to train a classifier for the clinical domain. However, it soon became evident that for the study of the two tasks, different types of input data would be required: The EHRs from endocrinology, which are comparatively extensive and are composed of linguistically distinct parts, formed an ideal basis for a terminological study; however, they contain few

spatial and temporal references, which were required to further develop methods in relation extraction. Therefore, as both project partners came to work on disjoint datasets, a synergy effect could not be reached.

9.3 Implications for Further Research

9.3.1 Terminology Management

The annotation of the clinical dataset showed that, in clinical records, the potential for term proliferation is huge. This provides further evidence that, to represent the terminology encountered in clinical usage, the exhaustive listing of all variants is not a viable approach, especially for low-resource languages like Belgian Dutch. However, the case study also showed that variation shows regularities depending on the semantic type of a concept, its combinatorial potential, and the register of the sublanguage it typically appears in. These regularities can be leveraged to enhance medical terminologies: For example, the individual concept entries could be enriched with details about possible variation processes at the type level. To derive the potential variants, the users of a terminology could follow a decision tree to select the processes that are most likely to apply, and generate the lexical forms based on transformational rules. This would enable a dynamic representation of terminology, where variation is accommodated without blowing up the knowledge base.

For instance, the case study showed that terms relating to ANATOMY are highly sensitive to the local context. If they co-occur with a DISORDER, they typically take the form of an adjectival modifier (e.g. *abdominale obesitas* 'abdominal obesity'). Together with PROCEDURES, though, where they form the object of investigation or therapy, noun forms more likely to appear (e.g. *palpatie van het abdomen* 'palpation of the abdomen'). This information could be integrated in a term base by applying a default rule for all concepts relating to ANATOMY: This rule would specify that these terms vary as a function of the local context, and that the variation processes tend to occur at

the morphological level. Furthermore, if the context involves a DISORDER, we are likely to encounter a derived adjective.

Term choices for PROCEDURES, on the other hand, are more prone to influences from the global context. In particular, reduction processes are very frequent in informal sublanguages, whereby the preferred type of reduction depends on the semantic composition of the term. In complex terms, constituents expressing the general action of examination tend to be left out (e.g. *schildklierfunctie<u>onderzoek</u>* 'thyroid function <u>examination</u>'– *schildklierfunctie* 'thyroid function'). On the other hand, if the head noun conveys semantically relevant details about the methodology, it will likely be preserved; however, the grammatical structure tends to be reduced (e.g. *rx thorax* 'x-ray thorax' instead of *rx <u>van de</u> thorax* 'x-ray <u>of the</u> thorax'). To represent this pattern in a term base, all concepts relating to PROCEDURES would need to be marked as sensitive to the global context, i.e. as showing variation as an effect of register switches. Moreover, a built-in list of general nouns relating to acts of examination would be required. Then, it could be specified that both the general examination nouns and function words might be left out, and that the order of the semantically relevant constituents could be permuted, depending on the sublanguage context.

9.3.2 Clinical NLP

The case study illustrated the significant impact that term variation could have on the automatic processing of clinical records. Among the terms encountered in our dataset, less than half of them were rated as *standard* (cf. Section 7.3.1). Evidently, an NLP application that does not take non-standard terms into account will miss vital information. Even within a single EHR, we encountered a number of sublanguages, which differ with regard to their terminological preferences. With an adequate scheme of operationalization, some of these preferences can be predicted with high accuracy. However, given the range of clinical specialties and documentation practices, the development and fine-tuning of a system to the specific properties of every clinical sublanguage is not a realistic scenario.

Still, clinical NLP could benefit from sublanguage analysis in order to handle term variation in a more efficient manner. For example, the case study showed that, within the context of a sublanguage, the formal properties of a term can be indicative of its semantic type. This phenomenon could be leveraged to support automatic term recognition from domain corpora: Term candidates acquired by automatic means could be assigned to semantic categories before being passed on to human experts for manual validation. The automatic pre-categorization of term variants would considerably reduce the manual workload required for the population of term bases in under-resourced languages.[15]

Likewise, such patterns could inform NER applications: As illustrated above, in informal contexts, terms relating to body parts and terms expressing their examination are used interchangeably, whereby the intended meaning is usually that of a PROCEDURE (e.g. *cor* 'heart' – *corauscultatie* 'heart auscultation'). This pattern could be implemented into a module for WSD: By adjusting the weights for the resolution of such forms depending on the sublanguage context, polysemous terms could be resolved more accurately.

9.4 Conclusion

Clinicians do not work with manufactured objects, but with human patients and their histories. Every patient presents an individual case; every observed phenomenon is the result of unique circumstances and differs in its own way from the prototypical understanding of the medical condition. Naturally, clinicians are reluctant to express themselves in standardized terms, let alone numerical codes, which might not do justice to the individual case. The complete standardization of clinical documentation in the near future is thus unlikely, and, from the point of view of the patient, also undesirable. For clinical NLP, the major challenge will thus remain the reconciliation of the hard codes required for computational processing, and the relative mess of human language. However, as exemplified by this thesis, there is some

[15] This idea is further elaborated in Grön, Bertels and Heylen (Forthcoming).

structure in this mess which can be attributed to the interactions between human perception, cognition and their expression. A deeper understanding of these interactions will make an essential contribution to further advances in clinical data reuse.

Bibliography

Adler-Milstein, Julia, and Ashish K. Jha. 2017. “HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption.” *Health Aff* 36 (8): 1416–22. https://doi.org/10.1377/hlthaff.2016.1651.

Afzal, Zubair, Ewoud Pons, Ning Kang, Miriam Sturkenboom, Martijn J. Schuemie, and Jan A. Kors. 2014. “ContextD: An Algorithm to Identify Contextual Properties of Medical Terms in a Dutch Clinical Corpus.” *BMC Bioinformatics* 15: 373. https://doi.org/10.1186/s12859-014-0373-3.

Albright, Daniel, Arrick Lanfranchi, Anwen Fredriksen, William F. Styler IV, Colin Warner, Jena D. Hwang, Jinho D. Choi, et al. 2013. “Towards Comprehensive Syntactic and Semantic Annotations of the Clinical Narrative.” *J Am Med Inform Assoc* 20: 922–930. https://doi.org/10.1136/amiajnl-2012-001317.

Allvin, Helen, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravicius, Martin Hassel, Dimitrios Kokkinakis, et al. 2011. “Characteristics of Finnish and Swedish Intensive Care Nursing Narratives: A Comparative Analysis to Support the Development of Clinical Language Technologies.” *J Biomed Sem* 2 (Suppl 3): S1. https://doi.org/10.1186/2041-1480-2-S3-S1.

Ananiadou, Sophia, and John McNaught. 2006. *Text Mining for Biology and Biomedicine*. Boston: Artech House.

Aramaki, Eiji, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2014. “Overview of the NTCIR-11 MedNLP-2 Task.” In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies*, edited by Noriko Kando, Hideo Joho, and Kazuaki Kishida, 147–54. Tokyo: National Institute of Informatics (NII).

———. 2016. “Overview of the NTCIR-12 MedNLPDoc Task.” In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, edited by Noriko Kando, Tetsuya Sakai, and Mark Sanderson, 71–75. Tokyo: National Institute of Informatics (NII).

Artstein, Ron. 2017. “Inter-Annotator Agreement.” In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, 297–314. Dordrecht: Springer.

Azevedo, Rafael F. de, Joao P. Santos Rodrigues, Mayara R. da Silva Reis, Claudia M. C. Moro, and Emerson C. Paraiso. 2018. “Temporal Tagging of Noisy Clinical Texts in Brazilian Portuguese. International Conference on Computational Processing of the Portuguese Language (PROPOR).” *LNCS* 11122: 231–41. https://doi.org/10.1007/978-3-319-99722-3.

Bansler, Jørgen P., Erling C. Havn, Kjeld Schmidt, Troels Mønsted, Helen Høgh Petersen, and Jesper Hastrup Svendsen. 2016. “Cooperative Epistemic Work in Medical Practice: An Analysis of Physicians’ Clinical Notes.” *CSCW* 25: 503–46. https://doi.org/10.1007/s10606-016-9261-x.

Belgisch Centrum voor Farmacotherapeutische Informatie. 2018. “Over Ons.” 2018. https://www.bcfi.be/nl/about.

———. 2019. “Gecommentarieerd Geneesmiddelenrepertorium.” 2019. https://www.bcfi.be/nl/chapters.

Bethard, Steven, Guergana K. Savova, Martha Palmer, and James Pustejovsky. 2018. “SemEval-2017 Task 12: Clinical TempEval.” In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, edited by Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, 565–72. Vancouver: Association for Computational Linguistics. https://doi.org/10.18653/v1/s17-2093.

Bodenreider, Olivier. 2004. “The Unified Medical Language System (UMLS): Integrating Biomedical Terminology.” *Nucleic Acids Res* 32 (Database Issue): D267–70. https://doi.org/10.1093/nar/gkh061.

Bodenreider, Olivier, Barry Smith, and Anita Burgun. 2004. “The Ontology-Epistemology Divide: A Case Study in Medical Terminology.” *Form Ontol Inf Syst*, 185–195.

Bousquet, Cédric, and Maria Zimina-Poirot. 2010. “The PERTOMed Project. Exploiting and Validating Terminological Resources of Comparable Russian-French-English Corpora within Pharmacovigilance.” In *Terminology in Everyday Life*, edited by Marcel Thelen and Frieda Steurs, 213–232. Amsterdam/ Philadelphia: John Benjamins.

Bowker, Lynne, and Shane Hawkins. 2007. “Variation in the Organization of Medical Terms: Exploring some Motivations for Term Choice.” *Terminology* 12 (1): 79–110. https://doi.org/10.1075/term.12.1.05bow.

Boytcheva, Svetla. 2012. “Multilingual Aspects of Information Extraction from Medical Texts in Bulgarian.” In *Multilingual Processing in Eastern and Southern EU Languages*, edited by Cristina Vertan and Walther von Hahn, 308–29. Newcastle upon Tyne: Cambridge Scholars Publishing.

Bozkurt, Selen, Francisco Gimenez, Elizabeth S. Burnside, Kemal H. Gulkesen, and Daniel L. Rubin. 2016. “Using Automatically Extracted Information from Mammography Reports for Decision-Support.” *J Biomed Inform* 62: 224–31. https://doi.org/10.1016/j.jbi.2016.07.001.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324.

Bretschneider, Claudia, and Sonja Zillner. 2015. “Semantic Splitting of German Medical Compounds.” *LNCS* 9302: 207–15. https://doi.org/10.1007/978-3-319-24033-6.

Bretschneider, Claudia, Sonja Zillner, and Matthias Hammon. 2013. “Identifying Pathological Findings in German Radiology Reports Using a Syntactico-Semantic Parsing Approach.” In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, edited by Kevin B. Cohen, Dina Demner-Fushman, Sophia Ananiadou, John Pestian, and Jun’ichi Tsujii, 27–35. Sofia: Association for Computational Linguistics.

Breydo, Eugene M., Julia T. Chu, and Alexander Turchin. 2008. “Identification of Inactive Medications in Narrative Medical Text.” *AMIA Annu Symp Proc*, 66–70.

Cabré Castellví, Maria T. 1999. *Terminology. Theory, Methods and Applications*. Edited by Juan C. Sager. Amsterdam/ Philadelphia: John Benjamins.

———. 2003. “Theories of Terminology. Their Description, Prescription and Explanation.” *Terminology* 2 (9): 163–99. https://doi.org/10.1075/term.9.2.03cab.

Cabré Castellví, Maria T., Carmen Bach, Rosa Estopà, Judit Feliu, Gemma Martínez, and Jorge Vivaldi. 2004. “The GENOMA-KB Project:

Towards the Integration of Concepts, Terms, Textual Corpora and Entities." In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, edited by Maria T. Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, 87–90. Lisbon: European Language Resources Association (ELRA).

Cabré Castellví, Maria T., Rosa E. Bagot, and Chelo Vargas-Sierra. 2012. "Neology in Specialized Communication." In *Neology in Specialized Communication*, edited by Maria T. Cabré Castellví, Rosa Estopà Bagot, and Chelo Vargas-Sierra, 1–8. Amsterdam/ Philadelphia: John Benjamins.

Cannon, Jay, and Susan Lucci. 2010. "Transcription and EHRs: Benefits of a Blended Approach." *J AHIMA* 81 (2): 36–40.

Carrell, David S., David Cronkite, Roy E. Palmer, Kathleen Saunders, David E. Gross, Elizabeth T. Masters, Timothy R. Hylan, and Michael Von Korff. 2015. "Using Natural Language Processing to Identify Problem Usage of Prescription Opioids." *Int J Med Inform* 84: 1057–64. https://doi.org/10.1016/j.ijmedinf.2015.09.002.

Carrell, David S., Scott Halgrim, Diem-Thy Tran, Diana S. M. Buist, Jessica Chubak, Wendy W. Chapman, and Guergana K. Savova. 2014. "Using Natural Language Processing to Improve Efficiency of Manual Chart Abstraction in Research: The Case of Breast Cancer Recurrence." *Am J Epidemiol* 179 (6): 749–58. https://doi.org/10.1093/aje/kwt441.

Chapman, Wendy W., Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries." *J Biomed Inform* 34: 301–10. https://doi.org/10.1006/jbin.2001.1029.

Chapman, Wendy W., David Chu, and John N. Dowling. 2007. "ConText: An Algorithm for Identifying Contextual Features from Clinical Text." In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, edited by Kevin B. Cohen, Dina Demner-Fushman, Carol Friedman, Lynette Hirschman, and John Pestian, 81–88. Prague: Association for Computational Linguistics.

Charrow, Vera, Jo A. Crandall, and Robert Charrow. 1982. "Characteristics and Functions of Legal Language." In *Sublanguages. Studies of*

Language in Restricted Semantic Domains, edited by Richard Kittredge and John Lehrberger, 175–90. Berlin/New York: De Gruyter.

Chen, Jinying, Abhyuday N. Jagannatha, Samah J. Fodeh, and Hong Yu. 2017. "Ranking Medical Terms to Support Expansion of Lay Language Resources for Patient Comprehension of Electronic Health Record Notes: Adapted Distant Supervision Approach." *JMIR Med Inform* 5 (4): e42. https://doi.org/10.2196/medinform.8531.

Chiaramello, Emma, Francesco Pinciroli, Alberico Bonalumi, Angelo Caroli, and Gabriella Tognola. 2016. "Use of 'Off-the-Shelf' Information Extraction Algorithms in Clinical Informatics: A Feasibility Study of MetaMap Annotation of Italian Medical Notes." *J Biomed Inform* 63: 22–32. https://doi.org/10.1016/j.jbi.2016.07.017.

Cimino, James J. 1998. "Desiderata for Controlled Medical Vocabularies in the Twenty-First Century." *Methods Inf Med* 37: 394–403.

Cohen, Kevin B., William A. Baumgartner, and Irina P. Temnikova. 2016. "SuperCAT: The (New and Improved) Corpus Analysis Toolkit." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, et al., 2784–88. Portorož: European Language Resources Association.

Cohen, Raphael. 2012. "Towards Understanding of Medical Hebrew." PhD diss., Ben-Gurion University of the Negev.

Cornet, Ronald, Carly Hill, and Nicolette De Keizer. 2017. "Comparison of Three English-to-Dutch Machine Translations of SNOMED CT Procedures." *Stud Health Technol Inform* 245: 848–52. https://doi.org/10.3233/978-1-61499-830-3-848.

Cruse, Alan D. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Daille, Béatrice. 2007. "Variations and Application-Oriented Terminology Engineering." In *Application-Driven Terminology Engineering*, edited by Fidelia Ibekwe-SanJuan, Anne Condamines, and Maria T. Cabré Castellví, 163–77. Amsterdam/ Philadelphia: John Benjamins.

———. 2018. *Term Variation in Specialised Corpora. Characterisation, Automatic Discovery and Applications*. Amsterdam/ Philadelphia: John

Benjamins.

Daille, Béatrice, Benoît Habert, Christian Jacquemin, and Jean Royauté. 1996. “Empirical Observation of Term Variations and Principles for Their Description.” *Terminology* 3 (2): 197–257. https://doi.org/10.1075/term.3.2.02dai.

Dalianis, Hercules. 2018. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Cham: Springer. https://doi.org/10.1007/978-3-319-78503-5.

De Block, Maggie, Jo Vandeurzen, Alda Greoli, Rudy Demotte, Didier Gosuin, Guy Vanhengel, Cécile Jodogne, and Antonios Antoniadis. 2019. “Actieplan E-Gezondheid 2019 -2021. Protocolakkoord.” https://www.ehealth.fgov.be/file/view/AWjHQ9zDgwvToiwBkf13?filename=Actieplan 2019-2021 e-Gezondheid_final.pdf.

Deléger, Louise, Magnus Merkel, and Pierre Zweigenbaum. 2009. “Translating Medical Terminologies through Word Alignment in Parallel Text Corpora.” *J Biomed Inform* 42: 692–701. https://doi.org/10.1016/j.jbi.2009.03.002.

della Volpe, Maddalena, Annibale Elia, and Francesca Esposito. 2018. “Semantic Predicates in the Business Language.” *CCIS* 811: 108–16. https://doi.org/10.1007/978-3-319-73420-0_9.

Demner-Fushman, Dina, and Noémie Elhadad. 2016. “Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing.” *Yearb Med Inform*, 224–33. https://doi.org/10.15265/iy-2016-017.

Deshors, Sandra C., and Stefan Th. Gries. 2016. “Profiling Verb Complementation Constructions across New Englishes.” *IJCL* 21 (2): 192–218. https://doi.org/10.1075/ijcl.21.2.03des.

Doing-Harris, Kristina, Yarden Livnat, and Stephane M. Meystre. 2015. “Automated Concept and Relationship Extraction for the Semi-Automated Ontology Management (SEAM) System.” *J Biomed Sem* 6: 15. https://doi.org/10.1186/s13326-015-0011-7.

Doing-Harris, Kristina, Olga Patterson, Sean Igo, and John F. Hurdle. 2013. “Document Sublanguage Clustering to Detect Medical Specialty in Cross-Institutional Clinical Texts.” *Proc ACM Int Workshop Data Text*

Min Biomed Inform, 9–12. https://doi.org/10.1145/2512089.2512101.

Drame, Khadim, Gayo Diallo, and Fleur Mougin. 2012. "Towards a Bilingual Alzheimer's Disease Terminology Acquisition Using a Parallel Corpus." *Stud Health Technol Inform* 180: 179–83. https://doi.org/10.3233/978-1-61499-101-4-179.

Dunham, George. 1986. "The Role of Syntax in the Sublanguage of Medical Diagnostic Statements." In *Analyzing Sublanguages in Restricted Domains. Sublanguage Description and Processing*, edited by Ralph Grishman and Richard Kittredge, 175–94. Hillsdale: Erlbaum.

Dziadek, Juliusz, Aron Henriksson, and Martin Duneld. 2017. "Improving Terminology Mapping in Clinical Text with Context-Sensitive Spelling Correction." *Stud Health Technol Inform* 235: 241–45. https://doi.org/10.3233/978-1-61499-753-5-241.

Elkin, Peter L., David A. Froehling, Dietlind L. Wahner-Roedler, Steven H. Brown, and Kent R. Bailey. 2012. "Comparison of Natural Language Processing Biosurveillance Methods for Identifying Influenza From Encounter Notes." *Ann Intern Med* 156: 11–18. https://doi.org/10.7326/0003-4819-156-1-201201030-00003.

Eriksson, Robert, Peter Bjødstrup Jensen, Sune Frankild, Lars Juhl Jensen, and Søren Brunak. 2013. "Dictionary Construction and Identification of Possible Adverse Drug Events in Danish Clinical Narrative Text." *J Am Med Inform Assoc* 20 (5): 947–53. https://doi.org/10.1136/amiajnl-2013-001708.

Evans, R. Scott. 2016. "Electronic Health Records: Then, Now, and in the Future." *Yearb Med Inform*, 25 Suppl 1: S48–61. https://doi.org/10.15265/iys-2016-s006.

Faber, Pamela. 2009. "The Cognitive Shift in Terminology and Specialized Translation." *MonTI* 1: 107–34.

———. , ed. 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin: De Gruyter Mouton.

Faber, Pamela, Pilar L. Araúz, Juan A. Prieto Velasco, and Arianne Reimerink. 2007. "Linking Images and Words: The Description of Specialized Concepts." *IJL* 20 (1): 39–65. https://doi.org/10.1093/ijl/ecl038.

Faber, Pamela, and Marie-Claude L'Homme. 2014. "Lexical Semantic

Approaches to Terminology." *Terminology* 20 (2). https://doi.org/https://doi.org/10.1075/term.20.2.

Faber, Pamela, and Pilar León-Araúz. 2016. "Specialized Knowledge Representation and the Parameterization of Context." *Front Psychol* 7 (February): 1–20. https://doi.org/10.3389/fpsyg.2016.00196.

Fan, Jung-Wei, Rashmi Prasad, Rommel M. Yabut, Richard M. Loomis, Daniel S. Zisook, John E. Mattison, and Yang Huang. 2011. "Part-of-Speech Tagging for Clinical Text: Wall or Bridge between Institutions?" *AMIA Annu Symp Proc*, 382–91.

Fan, Jung-Wei, Elly W. Yang, Min Jiang, Rashmi Prasad, Richard M. Loomis, Daniel S. Zisook, Joshua C. Denny, Hua Xu, and Yang Huang. 2013. "Syntactic Parsing of Clinical Text: Guideline and Corpus Development with Handling Ill-Formed Sentences." *J Am Med Inform Assoc* 20: 1168–77. https://doi.org/10.1136/amiajnl-2013-001810.

Felber, Helmut, ed. 1979. *Theory of Terminology and Terminological Lexicography*. Vienna/ New York: Springer.

Feldman, Keith, Nicholas Hazekamp, and Nitesh V. Chawla. 2016. "Mining the Clinical Narrative: All Text Are Not Equal." *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 271–80. https://doi.org/10.1109/ICHI.2016.37.

Ferraro, Jeffrey P., Hal Daumé Ill, Scott L. DuVall, Wendy W. Chapman, Henk Harkema, and Peter J. Haug. 2013. "Improving Performance of Natural Language Processing Part-of-Speech Tagging on Clinical Narratives through Domain Adaptation." *J Am Med Inform Assoc* 20: 931–39. https://doi.org/10.1136/amiajnl-2012-001453.

Fillmore, Charles J. 1992. "Frame Semantics." In *Linguistics in the Morning Calm*, edited by The Linguistic Society of Korea, 111–37. Seoul: Hanshin.

Finley, Gregory P., Serguei V. S. Pakhomov, Reed McEwan, and Genevieve B. Melton. 2016. "Towards Comprehensive Clinical Abbreviation Disambiguation Using Machine-Labeled Training Data." *AMIA Annu Symp Proc*, 560–69.

Fivez, Pieter, Simon Šuster, and Walter Daelemans. 2017. "Unsupervised Context-Sensitive Spelling Correction of Clinical Free-Text with Word and Character N-Gram Embeddings." In *BioNLP 2017*, edited by

Kevin B. Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Junichi Tsujii, 143–48. Vancouver: Association for Computational Linguistics.

Ford, Elizabeth, John A. Carroll, Helen E. Smith, Donia Scott, and Jackie A. Cassell. 2016. "Extracting Information from the Text of Electronic Medical Records to Improve Case Detection: A Systematic Review." *J Am Med Inform Assoc* 23: 1007–15. https://doi.org/10.1093/jamia/ocv180.

France, Francis R. 2011. "EHealth in Belgium, a New 'Secure' Federal Network: Role of Patients, Health Professions and Social Security Services." *Int J Med Inform* 80: e12–16. https://doi.org/10.1016/j.ijmedinf.2010.10.005.

Freixa, Judit. 2006. "Causes of Denominative Variation in Terminology: A Typology Proposal." *Terminology*, 51–77. https://doi.org/10.1075/term.12.1.04fre.

Friedman, Carol. 1986. "Automatic Structuring of Sublanguage Information: Application to Medical Narrative." In *Analyzing Sublanguages in Restricted Domains. Sublanguage Description and Processing*, edited by Ralph Grishman and Richard Kittredge, 85–102. Hillsdale: Erlbaum.

Friedman, Carol, Philip O. Alderson, John H. M. Austin, James J. Cimino, and Stephen B. Johnson. 1994. "A General Natural-Language Text Processor for Clinical Radiology." *J Am Med Inform Assoc* 1 (2): 161–74.

Friedman, Carol, Pauline Kra, and Andrey Rzhetsky. 2002. "Two Biomedical Sublanguages: A Description Based on the Theories of Zellig Harris." *J Biomed Inform* 35 (4): 222–35. https://doi.org/10.1016/S1532-0464(03)00012-1. https://doi.org/10.1016/S1532-0464(03)00012-1.

Friedman, Carol, Lyudmila Shagina, Yves Lussier, and George Hripcsak. 2004. "Automated Encoding of Clinical Documents Based on Natural Language Processing." *J Am Med Inform Assoc* 5: 392–402. https://doi.org/10.1197/jamia.M1552.The.

Friedman, Carol, Lyudmila Shagina, Socrates A. Socratous, and Xiao Zeng. 1996. "A WEB-Based Version of MedLEE: A Medical Language Extraction and Encoding System." In *Proc AMIA Annu Fall Symp*, edited by James J. Cimino, 938. Philadelphia: Hanley & Belfast.

Fung, Kin W., William T. Hole, Stuart J. Nelson, Suresh Srinivasan, Tammy Powell, and Laura Roth. 2005. “Integrating SNOMED CT into the UMLS: An Exploration of Different Views of Synonymy and Quality of Editing.” *J Am Med Inform Assoc* 12: 486–94. https://doi.org/10.1197/jamia.M1767.

Gardner, Reed M. 2016. “Clinical Information Systems – From Yesterday to Tomorrow.” *Yearb Med Inform* 25: S62-75.

Gaudin, François. 1993. *Pour Une Socioterminologie: Des Problèmes Pratiques Aux Pratiques Institutionnelles*. Rouen: Publications de l’Université de Rouen.

———. 2005. “La Socioterminologie.” *Langages* 157 (1): 80–92.

Goeuriot, Lorraine, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon. 2017. “CLEF 2017 EHealth Evaluation Lab Overview.” *LNCS* 10456: 291–303. https://doi.org/10.1007/978-3-319-65813-1_26.

Grigonyte, Gintare, Maria Kvist, Mats Wiren, Sumithra Velupillai, and Aron Henriksson. 2016. “Swedification Patterns of Latin and Greek Affixes in Clinical Text.” *Nor Jnl Ling* 39: 5–37. https://doi.org/10.1017/S0332586515000293.

Grön, Leonie, Ann Bertels, and Kris Heylen. 2018a. “Is Training Worth the Trouble? A PoS Tagging Experiment with Dutch Clinical Records.” In *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, 351–58. Rome: UniversItalia.

———. 2018b. “The Interplay of Form and Meaning in Complex Medical Terms: Evidence from a Clinical Corpus.” In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG)*, edited by Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan, and Miriam R. L. Petruck, 18–29. Santa Fe: Association for Computational Linguistics.

———. Forthcoming. “Leveraging Sublanguage Features for the Semantic Categorization of Clinical Terms.” In *Proceedings of the 18th Workshop on Biomedical Natural Language Processing*. Florence, Italy: Association for Computational Linguistics.

Groth Jensen, Lotte, and Claus Bossen. 2016. “Factors Affecting Physicians’

Use of a Dedicated Overview Interface in an Electronic Health Record: The Importance of Standard Information and Standard Documentation." *Int J Med Inform* 87: 44–53. https://doi.org/10.1016/j.ijmedinf.2015.12.009.

Grouin, Cyril, and Aurélie Névéol. 2014. "De-Identification of Clinical Notes in French: Towards a Protocol for Reference Corpus Development." *J Biomed Inform* 50: 151–61. https://doi.org/10.1016/j.jbi.2013.12.014.

Hamon, Thierry, and Natalia Grabar. 2014. "Tuning HeidelTime for Identifying Time Expressions in Clinical Texts in English and French." In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, edited by Sumithra Velupillai, Martin Duneld, Maria Kvist, Hercules Dalianis, Maria Skeppstedt, and Aron Henriksson, 101–5. Gothenburg: Association for Computational Linguistics.

Harkema, Henk, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. "ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports." *J Biomed Inform* 42: 839–51. https://doi.org/10.1016/j.jbi.2009.05.002.

Harris, Zellig S. 1982. "Discourse and Sublanguage." In *Sublanguages. Studies of Language in Restricted Semantic Domains*, edited by Richard Kittredge and John Lehrberger, 231–36. Berlin/New York.

———. 1991. *A Theory of Language and Information. A Mathematical Approach.* Oxford: Clarendon Press.

———. 2002. "The Structure of Science Information." *J Biomed Inform* 35 (4): 215–21. https://doi.org/10.1016/S1532-0464(03)00011-X.

Harris, Zellig S., and Paul Mattick. 1988. "Science Sublanguages and the Prospects for a Global Language of Science." *Ann Am Acad Pol Soc Sci* 495: 73–83. https://doi.org/10.1177/016344300022005001.

Hassanpour, Saeed, and Curtis P. Langlotz. 2016. "Information Extraction from Multi-Institutional Radiology Reports." *Artif Intell Med* 66: 29–39. https://doi.org/10.1016/j.artmed.2015.09.007.

Haverinen, Katri, Filip Ginter, Timo Viljanen, Veronika Laippala, and Tapio Salakoski. 2010. "Dependency-Based PropBanking of Clinical Finnish." In *Proceedings of the Fourth Linguistic Annotation Workshop*, edited by Nianwen Xue and Massimo Poesio, 137–41.

Uppsala: Association for Computational Linguistics.

Haverinen, Katri, Jenna Kanerva, Samuel Kohonen, Anna Missilä, Stina Ojala, Timo Viljanen, Veronika Laippala, and Filip Ginter. 2015. "The Finnish Proposition Bank." *Lang Resour Eval* 49: 907–26. https://doi.org/10.1007/s10579-015-9310-y.

He, Bin, Bin Dong, Yi Guan, Jinfeng Yang, Zhipeng Jiang, Qiubin Yu, Jianyi Cheng, and Chunyan Qu. 2017. "Building a Comprehensive Syntactic and Semantic Corpus of Chinese Clinical Texts." *J Biomed Inform* 69: 203–17. https://doi.org/10.1016/j.jbi.2017.04.006.

He, Zhe, Michael Halper, Yehoshua Perl, and Gai Elhanan. 2012. "Clinical Clarity versus Terminological Order – The Readiness of SNOMED CT Concept Descriptors for Primary Care." *MIXHS 12*, 1–6. https://doi.org/ 10.1145/2389672.2389674.

Hearst, Marti A. 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora Lexico-Syntactic for Hyponymy Patterns." In *COLING '92: Proceedings of the 14th Conference on Computational Linguistics*, 539–545. Stroudsburg: Association for Computational Linguistics. https://doi.org/10.1.1.36.701.

Hellrich, Johannes, Franz Matthies, Erik Faessler, and Udo Hahn. 2015. "Sharing Models and Tools for Processing German Clinical Texts." *Stud Health Technol Inform*, no. 210: 734–38. https://doi.org/10.3233/978-1-61499-512-8-734.

Henriksson, Aron, Mike Conway, Martin Duneld, and Wendy W. Chapman. 2013. "Identifying Synonymy between SNOMED Clinical Terms of Varying Length Using Distributional Analysis of Electronic Health Records." *AMIA Annu Symp Proc*, 600–609.

Henriksson, Aron, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius, and Martin Duneld. 2014. "Synonym Extraction and Abbreviation Expansion with Ensembles of Semantic Spaces." *J Biomed Sem* 5: 6. https://doi.org/10.1186/2041-1480-5-6.

Heyman, Geert, Ivan Vulić, and Marie-Francine Moens. 2018. "A Deep Learning Approach to Bilingual Lexicon Induction in the Biomedical Domain." *BMC Bioinformatics* 19 (1): 1–15. https://doi.org/10.1186/s12859-018-2245-8.

Ibáñez, Miguel S., and Joaquín G. Palacios. 2014. "Semantic

Characterization of Terms as a Trace of Terminological Dependency." In *Lexical Semantic Approaches to Terminology*, edited by Pamela Faber and Marie-Claude L'Homme, 171–97. Amsterdam/ Philadelphia: John Benjamins.

International Health Terminology Standards Development Organisation (IHTSDO). 2018. "SNOMED International July 2018 SNOMED CT International Edition Release Package."

———. 2019. "SNOMED CT Editorial Guide." 2019. https://confluence.ihtsdotools.org/display/EditorialGuide.

Iqbal, Ehtesham, Robbie Mallah, Daniel Rhodes, Honghan Wu, Alvin Romero, Nynn Chang, Olubanke Dzahini, et al. 2017. "ADEPt, a Semantically-Enriched Pipeline for Extracting Adverse Drug Events from Free-Text Electronic Health Records." *PLoS ONE* 12 (11): e0187121. https://doi.org/10.1371/journal.pone.0187121.

Isenius, Niklas, Maria Kvist, and Sumithra Velupillai. 2012. "Initial Results in the Development of SCAN: A Swedish Clinical Abbreviation Normalizer." In *CLEFeHealth2012 - The CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for EHealth Document Analysis*, edited by Pamela Forner, Jussi Karlgren, Christa Womser-Hacker, and Nicola Ferro. Rome: Conference and Labs of the Evaluation Forum.

Jiang, Min, Yang Huang, Jung-Wei Fan, Buzhou Tang, Joshua C. Denny, and Hua Xu. 2015. "Parsing Clinical Text: How Good Are the State-of-the-Art Parsers?" *BMC Med Inform Decis Mak* 15 (Suppl 1): S2. https://doi.org/10.1186/1472-6947-15-S1-S2.

Johnson, Stephen B., and Michael Gottfried. 1989. "Sublanguage Analysis as a Basis for a Controlled Medical Vocabulary." In *Proc Annu Symp Comput Appl Med Care*, edited by Lawrence C. Kingsland, 519–23. Washington/ Los Alamitos/ Brussels/ Tokyo: IEEE Computer Science Press.

Jonnalagadda, Siddhartha, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. 2012. "Enhancing Clinical Concept Extraction with Distributional Semantics." *J Biomed Inform* 45: 129–40. https://doi.org/10.1016/j.jbi.2011.10.007.

Juckett, David. 2012. "A Method for Determining the Number of Documents Needed for a Gold Standard Corpus." *J Biomed Inform* 45 (3): 460–70.

https://doi.org/10.1016/j.jbi.2011.12.010.

Kaipio, Johanna, Tinja Lääveri, Hannele Hyppönen, Suvi Vainiomäki, Jarmo Reponen, Andre Kushniruk, Elizabeth Borycki, and Jukka Vänskä. 2017. "Usability Problems Do Not Heal by Themselves: National Survey on Physicians' Experiences with EHRs in Finland." *Int J Med Inform* 97: 266–81. https://doi.org/10.1016/j.ijmedinf.2016.10.010.

Kara, Elif, Tatjana Zeen, Aleksandra Gabryszak, Klemens Budde, Danilo Schmidt, and Roland Roller. 2018. "A Domain-Adapted Dependency Parser for German Clinical Text." In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*, edited by Adrien Barbaresi, Hanno Biber, Friedrich Neubarth, and Rainer Osswald, 12–17. Vienna: Austrian Academy of Sciences.

Kate, Rohit. 2012. "Unsupervised Grammar Induction of Clinical Report Sublanguage." *J Biomed Sem* 3 (Suppl 3): S4. https://doi.org/10.1109/ICMLA.2011.150.

Kaufman, David R., Barbara Sheehan, Peter Stetson, Ashish R. Bhatt, Adele I. Field, Chirag Patel, and James M. Maisel. 2016. "Natural Language Processing-Enabled and Conventional Data Capture Methods for Input to Electronic Health Records: A Comparative Usability Study." *JMIR Med Inform* 28 (4): e35. https://doi.org/10.2196/medinform.5544.

Kennell, Timothy, James Willig, and James Cimino. 2018. "Clinical Informatics Researcher's Desiderata for the Data Content of the Next Generation Electronic Health Record." *Appl Clin Inform* 8 (4): 1159–72. https://doi.org/10.4338/aci-2017-06-r-0101.

Kerremans, Koen, Peter De Baer, and Rita Temmerman. 2010. "Competency-Based Job Descriptions and Termontography. The Case of Terminological Variation." In *Terminology in Everyday Life*, edited by Marcel Thelen and Frieda Steurs. Amsterdam/ Philadelphia: John Benjamins.

Kittredge, Richard. 2003. "Sublanguages and Controlled Languages." In *The Oxford Handbook of Computational Linguistics*, edited by Ruslan Mitkov. Oxford: Oxford University Press.

Knoll, Benjamin C., Genevieve B. Melton, Hongfang Liu, Hua Xu, and Serguei V. S. Pakhomov. 2016. "Using Synthetic Clinical Data to Train an HMM-Based POS Tagger." In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 252–55. Las

Vegas: IEEE. https://doi.org/10.1109/BHI.2016.7455882.

Koleck, Theresa A., Caitlin Dreisbach, Philip E. Bourne, and Suzanne Bakken. 2019. "Natural Language Processing of Symptoms Documented in Free-Text Narratives of Electronic Health Records: A Systematic Review." *J Am Med Inform Assoc* 26 (4): 364–79. https://doi.org/10.1093/jamia/ocy173.

Kreimeyer, Kory, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F. Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. "Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review." *J Biomed Inform* 73: 14–29. https://doi.org/10.1016/j.jbi.2017.07.012.

Kreuzthaler, Markus, and Stefan Schulz. 2015. "Detection of Sentence Boundaries and Abbreviations in Clinical Narratives." *BMC Med Inform Decis Mak* 15 (Suppl 2): S4.

Kushida, Clete A., Deborah A. Nichols, Rik Jadrnicek, Ric Miller, James K. Walsh, and Kara Griffin. 2012. "Strategies for De-Identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies." *Med Care* 50: S82-101. https://doi.org/10.1097/MLR.0b013e3182585355.

Kvist, Maria, and Sumithra Velupillai. 2014. "SCAN: A Swedish Clinical Abbreviation Normalizer. Further Development and Adaptation to Radiology." *LNCS* 8685: 62–73. https://doi.org/10.1007/978-3-319-11382-1_7

Lai, Kenneth H., Maxim Topaz, Foster R. Goss, and Li Zhou. 2018. "Automated Misspelling Detection and Correction in Clinical Free-Text Records." *J Biomed Inform* 55: 188–95. https://doi.org/10.1109/ICAIBD.2018.8396209.

Laippala, Veronika, Filip Ginter, Sampo Pyysalo, and Tapio Salakoski. 2009. "Towards Automated Processing of Clinical Finnish: Sublanguage Analysis and a Rule-Based Parser." *Int J Med Inform* 78: e7–12. https://doi.org/10.1016/j.ijmedinf.2009.02.005.

Laippala, Veronika, Timo Viljanen, Antti Airola, Jenna Kanerva, Sanna Salanterä, Tapio Salakoski, and Filip Ginter. 2014. "Statistical Parsing of Varieties of Clinical Finnish." *Artif Intell Med* 61: 131–36. https://doi.org/10.1016/j.artmed.2014.02.002.

Landis, J. Richard, and Gary G. Koch. 1977. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33 (1): 159–74.

Lang, Friedrich H. 1998. “Eugen Wüster - Zum 100. Geburtstag. Ein Elektrotechniker als Terminologe.” *Elektrotechnik und Informationstechnik* 115 (11): 625–28.

Litzelman, Debra K., Robert S. Dittus, Michael E. Miller, and William M. Tierney. 1993. “Requiring Physicians to Respond to Computerized Reminders Improves Their Compliance with Preventive Care Protocols.” *J Gen Intern Med* 8 (6): 311–17. https://doi.org/10.1007/bf02600144

Liu, Yue, Tao Ge, Kusum S. Mathews, Heng Ji, and Deborah L. McGuinness. 2015. “Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion.” In *Proceedings of BioNLP 15*, edited by Kevin B. Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Junichi Tsujii, 92–97. Beijing: Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3810.

Lohr, Christina, Sven Buechel, and Udo Hahn. 2018. “Sharing Copies of Synthetic Clinical Corpora without Physical Distribution - A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus.” In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, edited by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al., 1259–66. Miyazaki: European Language Resource Association.

Marshman, Elizabeth. 2014. “Enriching Terminology Resources with Knowledge-Rich Contexts. A Case Study.” In *Lexical Semantic Approaches to Terminology*, edited by Pamela Faber and Marie-Claude L’Homme, 225–249. Amsterdam/ Philadelphia: John Benjamins.

Martin-Sanchez, Fernando, and Karin Verspoor. 2014. “Big Data in Medicine Is Driving Big Changes.” *Yearb Med Inform* 91 (1): 14–20. https://doi.org/10.15265/IY-2014-0020.

Matsuki, Kazunaga, Victor Kuperman, and Julie A Van Dyke. 2016. “The Random Forests Statistical Technique: An Examination of Its Value for the Study of Reading.” *Sci Stud REad* 20 (1): 20–33. https://doi.org/10.1080/10888438.2015.1107073.The.

McCray, Alexa T., Anita Burgun, and Olivier Bodenreider. 2001. "Aggregating UMLS Semantic Types for Reducing Conceptual Complexity." *Stud Health Technol Inform* 84: 216–20. https://doi.org/10.3233/978-1-60750-928-8-216.

Mertens, Ingrid. 2018. "Belgische Release SNOMED CT." https://www.health.belgium.be/sites/default/files/uploads/fields/fpshealth_theme_file/nrc-be_03_belgische_release_snomedct_ingrid_mertens_20180328.pdf.

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

Meystre, Stephane M. 2015. "De-Identification of Unstructured Clinical Data for Patient Privacy Protection." In *Medical Data Privacy Handbook*, edited by Aris Gkoulalas-Divanis and Grigorios Loukides, 697–716. Cham: Springer. https://doi.org/10.1007/978-3-319-23633-9.

Meystre, Stephane M., Óscar Ferrández, Jeff Friedlin, Brett R. South, Shuying Shen, and Matthew H. Samore. 2014. "Text De-Identification for Privacy Protection: A Study of Its Impact on Clinical Text Information Content." *J Biomed Inform* 50: 142–50. https://doi.org/10.1016/j.jbi.2014.01.011.

Meystre, Stephane M., Christian Lovis, Thomas Bürkle, Gabriella Tognola, Andrius Budrionis, and Christoph U. Lehmann. 2017. "Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress." *Yearb Med Inform* 26: 38–52. https://doi.org/10.15265/IY-2017-007.

Meystre, Stephane M., Guergana K. Savova, Karin C. Kipper-Schuler, and John F. Hurdle. 2008. "Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research." *Methods Inf Med* 47 (S1): 128–44. https://doi.org/10.1055/s-0038-1638592.

Moon, Sungrim, Serguei V. S. Pakhomov, Nathan Liu, James O. Ryan, and Genevieve B. Melton. 2014. "A Sense Inventory for Clinical Abbreviations and Acronyms Created Using Clinical Notes and Medical Dictionary Resources." *J Am Med Inform Assoc* 21: 299–307.

https://doi.org/10.1136/amiajnl-2012-001506.

National ICT Instituut in de Zorg (Nictiz). 2018. “SNOMED CT - Netherlands Edition 31 March 2018.”

National Library of Medicine. 2018. “Semantic Types and Groups.” 2018. https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml.

———. 2019a. “UMLS Release File Archives: 2004-2018AB. 2018AA UMLS Release Files.” 2019. https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html#2018AA.

———. 2019b. “UMLS Terminology Services.” https://uts.nlm.nih.gov/home.html.

Neamatullah, Ishna, Margaret M. Douglass, Li-Wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. 2008. “Automated De-Identification of Free-Text Medical Records.” *BMC Med Inform Decis Mak* 8: 1–16. https://doi.org/10.1186/1472-6947-8-32.

Névéol, Aurélie, Kevin B. Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, et al. 2016. “Clinical Information Extraction at the CLEF EHealth Evaluation Lab 2016.” *CEUR Workshop Proc* 1609: 28–42.

Névéol, Aurélie, Hercules Dalianis, Sumithra Velupillai, Guergana K. Savova, and Pierre Zweigenbaum. 2018a. “Clinical Natural Language Processing in Languages Other than English: Opportunities and Challenges.” *J Biomed Sem* 9: 12.

Névéol, Aurélie, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, László Pelikan, Lionel Ramadier, Grégoire Rey, and Pierre Zweigenbaum. 2018b. “CLEF EHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian.” *CEUR Workshop Proc*: 2125.

Névéol, Aurélie, and Pierre Zweigenbaum. 2018. “Expanding the Diversity of Texts and Applications: Findings from the Section on Clinical Natural Language Processing of the International Medical Informatics Association Yearbook.” *Yearb Med Inform*, 193–98. https://doi.org/10.1055/s-0038-1667080.

Nguyen, Hoang, and Jon Patrick. 2016. “Text Mining in Clinical Domain:

Dealing with Noise." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, edited by Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, 549–58. San Francisco: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939720.

Oleynik, Michel, Percy Nohama, Pindaro Secco Cancian, and Stefan Schulz. 2010. "Performance Analysis of a POS Tagger Applied to Discharge Summaries in Portuguese." *Stud Health Technol Inform* 160: 959–63. https://doi.org/10.3233/978-1-60750-588-4-959.

Pakhomov, Serguei V. S., Anni Coden, and Christopher G. Chute. 2006. "Developing a Corpus of Clinical Notes Manually Annotated for Part-of-Speech." *Int J Med Inform* 75 (6): 418–29. https://doi.org/10.1016/j.ijmedinf.2005.08.006.

Patterson, Olga, and John F. Hurdle. 2011. "Document Clustering of Clinical Narratives: A Systematic Study of Clinical Sublanguages." *AMIA Annu Symp Proc*, 1099–1107.

Patterson, Olga, Sean Igo, and John F. Hurdle. 2010. "Automatic Acquisition of Sublanguage Semantic Schema: Towards the Word Sense Disambiguation of Clinical Narratives." *AMIA Annu Symp Proc*, 612–16.

Pérez, Alicia, Rebecka Weegar, Arantza Casillas, Koldo Gojenola, Maite Oronoz, and Hercules Dalianis. 2017. "Semi-Supervised Medical Entity Recognition : A Study on Spanish and Swedish Clinical Corpora." *J Biomed Inform* 71: 16–30. https://doi.org/10.1016/j.jbi.2017.05.009.

Perotte, Adler, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. "Diagnosis Code Assignment: Models and Evaluation Metrics." *J Am Med Inform Assoc* 21: 231–37. https://doi.org/10.1136/amiajnl-2013-002159.

Peterson, Kevin J., and Hongfang Liu. 2018. "The Sublanguage of Clinical Problem Lists: A Corpus Analysis." *AMIA Annu Symp Proc*, 1451–60.

Pradhan, Sameer, Noémie Elhadad, Brett R. South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana K. Savova. 2015. "Evaluating the State of the Art in Disorder Recognition and Normalization of the Clinical Narrative." *J Am Med*

Inform Assoc 22: 143–54. https://doi.org/10.1136/amiajnl-2013-002544.

Prieto Velasco, Juan A., and Maribel Tercedor Sánchez. 2014. "The Embodied Nature of Medical Concepts: Image Schemas and Language for Pain." *Cogn Proc* 15 (3): 283–96. https://doi.org/10.1007/s10339-013-0594-9.

Rector, Alan L. 1999. "Clinical Terminology: Why Is It so Hard?" *Methods Inf Med* 38: 239–52.

Richter-Pechanski, Phillip, Stefan Riezler, and Christoph Dieterich. 2018. "De-Identification of German Medical Admission Notes." *Stud Health Technol Inform* 253: 165–69. https://doi.org/10.3233/978-1-61499-896-9-165.

Roldán Vendrell, Mercedes, and Jesús Fernández-Domínguez. 2012. "Emergent Neologisms and Lexical Gaps in Specialised Languages." In *Neology in Specialized Communication*, edited by Maria T. Cabré Castellví, Rosa Estopà Bagot, and Chelo Vargas-Sierra, 9–26. Amsterdam/ Philadelphia: John Benjamins.

Roller, Roland, Hans Uszkoreit, Feiyu Xu, Laura Seiffe, Michael Mikhailov, Oliver Staeck, Klemens Budde, Fabian Halleck, and Danilo Schmidt. 2016. "A Fine-Grained Corpus Annotation Schema of German Nephrology Records." In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, edited by Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, 69–77. Osaka: The COLING 2016 Organizing Committee.

Rosenbloom, S. Trent, Joshua C. Denny, Hua Xu, Nancy Lorenzi, William W. Stead, and Kevin B. Johnson. 2011. "Data from Clinical Notes: A Perspective on the Tension between Structure and Flexible Documentation." *J Am Med Inform Assoc* 18: 181–86. https://doi.org/10.1136/jamia.2010.007237.

Rubio-López, Ignacio, Roberto Costumero, and Héctor Ambit. 2017. "Acronym Disambiguation in Spanish Electronic Health Narratives Using Machine Learning Techniques." *Stud Health Technol Inform* 235: 251–55. https://doi.org/10.3233/978-1-61499-753-5-251.

Ruud, Kari L., Matthew G. Johnson, Juliette T. Liesinger, Carrie A. Grafft, and James M. Naessens. 2010. "Automated Detection of Follow-Up Appointments Using Text Mining of Discharge Records." *Int J Qual*

Health Care 22 (3): 229–35. https://doi.org/10.1093/intqhc/mzq012.

Safran, Charles, Meryl Bloomrosen, Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C. Tang, and Don E. Detmer. 2007. "Toward a National Framework for the Secondary Use of Health Data: An AMIA White Paper." *J Am Med Inform Assoc* 14 (1): 1–9. https://doi.org/10.1197/jamia.M2273.Introduction.

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

Sager, Naomi, Carol Friedman, and Margaret S. Lyman. 1987. *Medical Language Processing: Computer Management of Narrative Data*. Boston: Addison-Wesley Longman.

Sager, Naomi, Margaret S. Lyman, Christine Bucknall, Ngo Nhan, and Leo J. Tick. 1994. "Natural Language Processing and the Representation of Clinical Data." *J Am Med Inform Assoc* 1 (2): 142–60. https://doi.org/10.1136/jamia.1994.95236145.

Sambre, Paul, and Cornelia Wermuth. 2010. "Instrumentality in Cognitive Concept Modelling." In *Terminology in Everyday Life*, edited by Marcel Thelen and Frieda Steurs, 233–54. Amsterdam/ Philadelphia: John Benjamins.

Savova, Guergana K., Jin Fan, Zi Ye, Sean P. Murphy, Jiaping Zheng, Christopher G. Chute, and Iftikhar J. Kullo. 2010a. "Discovering Peripheral Arterial Disease Cases from Radiology Notes Using Natural Language Processing." In *AMIA Annu Symp Proc*, 722–26. Association for Computational Linguistics.

Savova, Guergana K., James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010b. "Mayo Clinical Text Analysis and Knowledge Extraction System (CTAKES): Architecture, Component Evaluation and Applications." *J Am Med Inform Assoc* 17: 507–13. https://doi.org/10.1136/jamia.2009.001560.

Savova, Guergana K., Sameer Pradhan, Martha Palmer, Will Styler, Noémie Elhadad, and Wendy W. Chapman. 2017. "Annotating the Clinical Text – MiPACQ, ShARe, SHARPn and THYME Corpora." In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, 1357–78. Dordrecht: Springer.

Scheurwegs, Elyne, Kim Luyckx, Léon Luyten, Bart Goethals, and Walter Daelemans. 2017. “Assigning Clinical Codes with Data-Driven Concept Representation on Dutch Clinical Free Text.” *J Biomed Inform* 69: 118–27. https://doi.org/10.1016/j.jbi.2017.04.007.

Schulz, Stefan, Johannes Bernhardt-Melischnig, Markus Kreuzthaler, Philipp Daumke, and Martin Boeker. 2013. “Machine vs. Human Translation of SNOMED CT Terms.” *Stud Health Technol Inform* 192 (1–2): 581–84. https://doi.org/10.3233/978-1-61499-289-9-581.

Siklósi, Borbála, and Attila Novák. 2013. “Detection and Expansion of Abbreviations in Hungarian Clinical Notes.” Edited by Félix Castro, Alexander Gelbkuh, and Miguel Gonzáles. *LNCS* 8265: 318–28. https://doi.org/https://doi.org/10.1007/978-3-642-45114-0_26.

Skeppstedt, Maria, and Aron Henriksson. 2013. “Vocabulary Expansion by Semantic Extraction of Medical Terms.” In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*, edited by Fabio Rinaldi and Jin-Dong Kim, 63–68. Tokyo: Database Center for Life Science.

Skeppstedt, Maria, Maria Kvist, Gunnar H. Nilsson, and Hercules Dalianis. 2014. “Automatic Recognition of Disorders, Findings, Pharmaceuticals and Body Structures from Clinical Text: An Annotation and Machine Learning Study.” *J Biomed Sem* 49: 148–58. https://doi.org/10.1016/j.jbi.2014.01.012.

Skevofilakas, Marios, Konstantia Zarkogianni, Basil G. Karamanos, and Konstantina S. Nikita. 2010. “A Hybrid Decision Support System for the Risk Assessment of Retinopathy Development as a Long Term Complication of Type 1 Diabetes Mellitus.” In *Conf Proc IEEE Eng Med Biol Soc*, 6713–16. Piscataway: IEEE. https://doi.org/10.1109/IEMBS.2010.5626245.

SNOMED International. 2019. “SNOMED CT July 2017 International Edition - SNOMED International Release Notes.” 2019. https://confluence.ihtsdotools.org/display/RMT/SNOMED+CT+July+2017+International+Edition+-+SNOMED+International+Release+notes.

Spyns, Peter. 1996. “A Dutch Medical Language Processor.” *Int J Biomed Comput* 41: 181–205. https://doi.org/10.1016/0020-7101(96)01198-1

———. 2000. *Natural Language Processing in Medicine. Design, Implementation and Evaluation of an Analyser for Dutch*. Leuven:

Leuven University Press.

St-Maurice, Justin, Ming-Han Kuo, and Phil Gooch. 2013. "A Proof of Concept for Assessing Emergency Room Use with Primary Care Data and Natural Language Processing." *Methods Inf Med* 52 (1): 33–42.

Strötgen, Jannik, and Michael Gertz. 2010. "HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions" In *Proceedings of the 5th International Workshop on Semantic Evaluation*, edited by Katrin Erk and Carlo Strapparava, 321–24. Uppsala: Association for Computational Linguistics.

Sun, Weiyi, Anna Rumshisky, and Özlem Uzuner. 2013. "Temporal Reasoning over Clinical Text: The State of the Art." *J Am Med Inform Assoc* 20: 814–19. https://doi.org/10.1136/amiajnl-2013-001760.

Tao, Carson, Michele Filannino, and Özlem Uzuner. 2017. "Prescription Extraction Using CRFs and Word Embeddings." *J Biomed Sem* 72: 60–66. https://doi.org/10.1016/j.jbi.2017.07.002.

Tavares, Jorge, and Tiago Oliveira. 2017. "Electronic Health Record Portal Adoption: A Cross Country Analysis." *BMC Med Inform Decis Mak* 17: 97. https://doi.org/10.1186/s12911-017-0482-9.

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

Temnikova, Irina P., William A. Baumgartner, Negacy D. Hailu, Ivelina Nikolova, Tony. McEnery, Adam Kilgarriff, Galia Angelova, and Kevin B. Cohen. 2014. "Sublanguage Corpus Analysis Toolkit: A Tool for Assessing the Representativeness and Sublanguage Characteristics of Corpora." *LREC Int Conf Lang Resour Eval*, 1714–18.

Temnikova, Irina P., Ivelina Nikolova, William A. Baumgartner, Galia Angelova, and Kevin B. Cohen. 2013. "Closure Properties of Bulgarian Clinical Text." *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, 667–75.

Tercedor Sánchez, Maribel. 2011. "The Cognitive Dynamics of Terminological Variation." *Terminology* 17 (2): 181–97. https://doi.org/10.1075/term.17.2.01ter.

Thomas, Cecilia E., Peter B. Jensen, Thomas Werge, and Søren Brunak. 2014. "Negation Scope and Spelling Variation for Text-Mining of

Danish Electronic Patient Records." In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, edited by Sumithra Velupillai, Martin Duneld, Maria Kvist, Hercules Dalianis, Maria Skeppstedt, and Aron Henriksson, 64–68. Gothenburg: Association for Computational Linguistics.

Thompson, Paul, and Sophia Ananiadou. 2018. "HYPHEN: A Flexible, Hybrid Method to Map Phenotype Concept Mentions to Terminological Resources." *Terminology* 24 (1): 91–121. https://doi.org/10.1075/term.00015.tho

Turchin, Alexander, Holly I. Wheeler, Matthew Labreche, Julia T. Chu, Merri L. Pendergrass, and Jonathan S. Einbinder. 2008. "Identification of Documented Medication Non-Adherence in Physician Notes." *AMIA Annu Symp Proc*, 732–36.

Vagelatos, Aristides, Elena Mantzari, Mavina Pantazara, Christos Tsalidis, and Chryssoula Kalamara. 2011. "Developing Tools and Resources for the Biomedical Domain of the Greek Language." *Health Inform J* 17 (2): 127–39. https://doi.org/10.1177/1460458211405007.

Velupillai, Sumithra, Danielle Mowery, Brett R. South, Maria Kvist, and Hercules Dalianis. 2015. "Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis." *Yearb Med Inform* 10: 183–93. https://doi.org/10.15265/iy-2015-009.

Wang, Xiaoyan, George Hripcsak, Marianthi Markatou, and Carol Friedman. 2009. "Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study." *J Am Med Inform Assoc* 16: 328–37. https://doi.org/10.1197/jamia.M3028.

Wang, Yan, Serguei V. S. Pakhomov, James O. Ryan, and Genevieve B. Melton. 2015. "Domain Adaption of Parsing for Operative Notes." *J Biomed Inform* 54: 1–9. https://doi.org/10.1016/j.jbi.2015.01.016.

Wang, Yanshan, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, et al. 2018. "Clinical Information Extraction Applications: A Literature Review." *J Biomed Inform* 77: 34–49. https://doi.org/10.1016/j.jbi.2017.11.011.

Wermter, Joachim, and Udo Hahn. 2004. "Really, Is Medical Sublanguage That Different? Experimental Counter-Evidence from Tagging Medical and Newspaper Corpora." *Stud Health Technol Inform* 107: 560–64.

Wermuth, Maria-Cornelia, and Heidi Verplaetse. 2019. “Medical Terminology in the Western World. Current Situation.” In *Handbook of Terminology. Terminology in the Arab World*, edited by Abied Alsulaiman and Ahmed Allaithy, 110–37. Amsterdam/ Philadelphia: John Benjamins.

Workman, T. Elizabeth, Yijun Shao, Guy Divita, and Qing Zeng-Treitler. 2019. “An Efficient Prototype Method to Identify and Correct Misspellings in Clinical Text.” *BMC Res Notes* 12: 42. https://doi.org/10.1186/s13104-019-4073-y.

Wu, Stephen, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. “Negation’s Not Solved: Generalizability versus Optimizability in Clinical Natural Language Processing.” *PLoS ONE* 9 (11): e112774. https://doi.org/10.1371/journal.pone.0112774.

Wu, Yonghui, Jianbo Lei, Wei Qi Wei, Buzhou Tang, Joshua C. Denny, S. Trent Rosenbloom, Randolph A. Miller, Dario A. Giuse, Kai Zheng, and Hua Xu. 2013. “Analyzing Differences Between Chinese and English Clinical Text: A Cross-Institution Comparison of Discharge Summaries in Two Languages.” *Stud Health Technol Inform* 192: 662–66. https://doi.org/10.3233/978-1-61499-289-9-662.

Wüster, Eugen. 1968. *The Machine Tool: An Interlingual Dictionary of Basic Concepts*. London: Technical Press.

———. 1979. *Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie*. Vienna/ New York: Springer.

Xu, Hua, Min Jiang, Matt Oetjens, Erica A. Bowton, Andrea H. Ramirez, Janina M. Jeff, Melissa A. Basford, et al. 2011a. “Facilitating Pharmacogenetic Studies Using Electronic Health Records and Natural-Language Processing: A Case Study of Warfarin.” *J Am Med Inform Assoc* 18: 387–91. https://doi.org/10.1136/amiajnl-2011-000208.

Xu, Hua, Samir A. Rahman, Yanxin Lu, Joshua C. Denny, and Son Doan. 2011b. “Applying Semantic-Based Probabilistic Context-Free Grammar to Medical Language Processing - A Preliminary Study on Parsing Medication Sentences.” *J Biomed Inform* 44: 1068–75. https://doi.org/10.1016/j.jbi.2011.08.009.

Xu, Peng, and Frederick Jelinek. 2004. “Random Forests in Language

Modeling." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 325–32. Barcelona: Association for Computational Linguistics.

Xu, Yan, Luoxin Chen, Junsheng Wei, Sophia Ananiadou, Yubo Fan, Yi Qian, Eric I.Chao Chang, and Junichi Tsujii. 2015. "Bilingual Term Alignment from Comparable Corpora in English Discharge Summary and Chinese Discharge Summary." *BMC Bioinformatics* 16: 1–10. https://doi.org/10.1186/s12859-015-0606-0.

Zeng, Qing T., Doug Redd, Guy Divita, Samah Jarad, Cynthia Brandt, and Jonathan R. Nebeker. 2011. "Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes." *J Health Med Informat* S3: 1–9. https://doi.org/10.4172/2157-7420.s3-001.

Zeng, Qing T., and Tony Tse. 2006. "Exploring and Developing Consumer Health Vocabularies." *J Am Med Inform Assoc* 13: 24–29. https://doi.org/10.1197/jamia.M1761.

Zhang, Rui, Jialin Liu, Yong Huang, Miye Wang, Qingke Shi, Jun Chen, and Zhi Zeng. 2017. "Enriching the International Clinical Nomenclature with Chinese Daily Used Synonyms and Concept Recognition in Physician Notes." *BMC Med Inform Decis Mak* 17: 54. https://doi.org/10.1186/s12911-017-0455-z.

Zhang, Shaodian, Tian Kang, Xingting Zhang, Dong Wen, Noémie Elhadad, and Jianbo Lei. 2016. "Speculation Detection for Chinese Clinical Notes: Impacts of Word Segmentation and Embedding Models." *J Biomed Inform* 60: 334–41. https://doi.org/10.1016/j.jbi.2016.02.011.

Zhao, Yiqing, Nooshin J. Fesharaki, Hongfang Liu, and Jake Luo. 2018. "Using Data-Driven Sublanguage Pattern Mining to Induce Knowledge Models: Application in Medical Image Reports Knowledge Representation Philip Payne." *BMC Med Inform Decis Mak* 18: 61. https://doi.org/10.1186/s12911-018-0645-3.