The Interplay of Form and Meaning in Complex Medical Terms: Evidence from a Clinical Corpus

Leonie Grön
KU Leuven, Belgium
leonie.gron@kuleuven.be

Ann Bertels
KU Leuven, Belgium
ann.bertels@kuleuven.be

Kris Heylen
KU Leuven, Belgium
kris.heylen@kuleuven.be

Abstract

We conduct a corpus study to investigate the structure of multi-word expressions (MWEs) in the clinical domain. Based on an existing medical taxonomy, we develop an annotation scheme and label a sample of MWEs from a Dutch corpus with semantic and grammatical features. The analysis of the annotated data shows that the formal structure of clinical MWEs correlates with their conceptual properties. The insights gained from this study could inform the design of Natural Language Processing (NLP) systems for clinical writing, but also for other specialized genres.

1 Introduction

Meaning, in everyday language, is created by the interaction of words in context, not simply by the words themselves. Word combinations are governed by grammatical rules and semantic constraints. However, some sequences show a distinctive *idiomaticity*: They either occur with an outstanding frequency, defy grammatical rules or convey a meaning that goes beyond the sum of their parts (Baldwin and Kim, 2010). Over the past decades, a number of frameworks has been proposed to approach such sequences from different theoretical angles. Consequently, they have been variably referred to as collocations, lexical bundles or multi-word expressions (MWEs), to name but a few terms. MWEs attract continuous scientific attention, not only due to their prevalence in the lexicon (Jackendoff, 1997), but also because they are a crucial factor in the performance of NLP: A system that cannot handle idiomatic expressions will at least miss semantic nuances, or fail to interpret the input completely.

Similarly, in specialized discourses, units of information are rarely encoded by a single term; instead, the majority of domain-specific concepts is referred to by MWEs, in particular, complex noun phrases (Daille, 1994; De Hertog & Heylen, 2012). Structured knowledge sources, which not only list the domain-specific vocabulary, but also model taxonomic relations, can serve to predict co-occurrences at the conceptual level. For instance, SNOMED CT (SNOMED International, 2018a) is a systematic medical terminology, which is primarily used to assign conceptual codes to electronic health records (EHRs). In SNOMED CT, each term is linked to a concept that belongs to a semantic category (e.g. *findings*, such as a symptom or a disease, and *procedures*, such as a method of investigation or therapy). Depending on the semantic category, each concept can be modified by a fixed set of qualifiers: A finding can be classified with regard to its severity, whereas a procedure can be specified with regard to the device used. However, this taxonomy provides little guidance about how the combination of the terms appears in practice. For instance, the MWE *obese abdomen* comprises two atomic concepts, the primary finding (*obesity*) and the anatomical site (*abdomen*). In Dutch, this observation can be expressed by a pre-modified noun phrase (*abdominale obesitas* 'abdominal obesity'), an attributive construction (*abdomen obees* 'abdomen obese'), a prepositional phrase (*obesitas thv abdomen*, which is short for *obesitas ter hoogte van abdomen* 'obesity at the abdomen'), or a reduced version thereof (*obesitas*

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 18–29
Santa Fe, New Mexico, USA, August 25-26, 2018.

abdomen 'obesity abdomen'). Given the productivity of morpho-syntactic processes, the exhaustive listing of all possible forms is obviously impractical.

According to cognitive theories of terminology, though, morpho-syntactic alternations in specialized terms are not arbitrary. Instead, variation is functional and can serve to convey different semantic nuances (Bowker & Hawkins, 2006; Faber et al., 2010). Likewise, constructional approaches to language posit that the pairing of syntactic structures with semantic components can create meaning in itself (Goldberg, 1995; Goldberg 2003). It is thus conceivable that habitual combinations of medical concepts pattern with particular structures at the formal level. If this is the case, such patterns could be leveraged to improve applications for clinical language processing, such as the automatic mapping of terms to ontological concepts.

To investigate this hypothesis, we conduct a corpus study, whereby we analyze sample of MWEs relating to different types of medical concepts. We develop an annotation scheme that captures grammatical and syntactic features of the MWE, and the semantic properties of the constituents. Using this annotation scheme, we analyze the interplay of semantic properties with the syntactic and grammatical structure of the MWEs, and their degree of lexicalization.

The remainder of this paper is structured as follows: In Section 2, we give an overview of related research to sketch the background of this work. In Section 3, we present our methodology: After introducing our dataset, we outline our approach for the identification, annotation and evaluation of the MWEs. Following the summary of the results (Section 4), we discuss the main findings of our study. Finally, in Section 5, we conclude with the implications of our work for future research on the processing of MWEs.

2 Background

As the scope of phraseological studies has widened considerably over the past decades, the names and definitions of MWEs have proliferated. In this paper, we adopt the definition proposed by Baldwin and Kim (2010), who define MWEs as expressions that consist of multiple lexical items, and that are marked by idiomaticities at the lexical, syntactic, semantic, pragmatic and/or statistical level. As Ramisch (2015) points out, this definition does not make any restrictions with regard to the granularity of the lexical items as they appear in the surface form; consequently, even orthographically joint forms, such as compound nouns can be considered MWEs. Likewise, this definition does not confine the notion of idiomatic behavior to one level of linguistic analysis. Traditionally, semantic opacity has been considered a main property of MWEs (Choueka, 1988; Fillmore et al., 1988; Smadja, 1993); conversely, this approach allows to treat compositionality as a continuum, and include semantically opaque and fully compositional MWEs alike.

With regard to languages for special purposes (LSPs), the central role of MWEs has long been acknowledged: Schulze and Römer (2008) emphasize that phraseological items are inseparably intertwined with the respective domain, as certain linguistic structures are instrumental to convey specialized meanings. The relationship between LSPs and their domain is thus dialectal in nature: While the language is a central constituent of the domain, the domain shapes the communicative needs. These needs shape syntactic and morphological patterns, which become consolidated through repeated usage.

Faber and Léon-Araùz (2016) propose a taxonomy of phraseal context factors, which is based on scope (local vs. global) and the dimension of linguistic analysis (syntactic, semantic or pragmatic). In this framework, MWEs are interpreted as a form of local context, whereby recurrent contexts can be modelled by syntagmatic grammatical patterns. Crucially, they observe a mutual attraction between syntactic and semantic structures; therefore, grammatical sequences and semantic relations should not be studied in isolation.

Similar patterns have been observed with nominal compounds. In general, the semantic relation between the constituents is indeterminable, as the underlying relation has been deleted or submerged as part of the compounding process (Levi, 1978). The full meaning can thus only be inferred through domain knowledge (Cabezas-García & San Martín, 2017). However, ten Hacken (2015) notes that, in biomedical compounds, the grammatical properties of the headword can give an indication of the relationship with the other constituent (i.e. the non-head).

Thus far, however, the analysis of medical MWEs has mostly focused on the biomedical domain, both with regard to phraseal expressions (e.g. Léon & Divasson, 2006; Laso & John, 2013; Lossio-Ventura,

Jonquet, Roche & Teisseire, 2016) and compounds (e.g. Cabezas-García & San Martín, 2017; Yadav et al., 2017). Meanwhile, the usage of MWEs in clinical practice has received relatively little attention. One obvious reason for this asymmetry is that clinical datasets are notoriously difficult to obtain. In English, a number of datasets has been labelled with medical identifiers (i.e. codes from a structured knowledge source such as SNOMED CT) and made available in the context of shared tasks, such as the i2b2 challenges (i2b2 National Center for Biomedical Computing, 2018) and the CLEF eHealth evaluation labs (CLEF eHealth, 2018). In addition, a number of corpora has also been annotated with semantic relations, PoS tags and, to varying depths, syntactic structure (e.g. Pakhomov, Coden & Chute, 2006; Roberts et al., 2009; Uzuner, Solti & Cadag, 2010; Uzuner, South, Shen & DuVall, 2011; Sum, Rumshisky & Uzuner, 2013; Albright et al., 2013; Styler et al., 2014; Savkov et al., 2016). However, for languages other than English, the distribution of clinical datasets is severely limited by the even more complex privacy regulations outside of the United States (cf. Névéol et al., 2018 for a comprehensive review). Still, an increasing number of – albeit not shareable – datasets has been annotated with both semantic and syntactic information, e.g. for Portuguese (Oleynik et al., 2010), Polish (Marciniak & Mykowiecka, 2011), Finnish (Haverinen et al., 2011; Laippala et al. 2014), Spanish (Costumero et al., 2014; Oronoz et al., 2015) and French (Deléger et al. 2017). In Dutch, clinical corpora have been labelled with features relating to negation, temporality and experiencer (Afzal et al., 2014) and codes from ICD-9 (Scheurwegs et al., 2017). However, we are not aware of any annotation projects that cover syntactic or grammatical properties of clinical Dutch; this paper thus presents a first attempt to fill this gap.

3 Methods

3.1 Corpus Structure

Our analysis is based on a corpus of Dutch EHRs provided by the department of endocrinology of a Belgian hospital. In total, this corpus consists of 14,999 documents and covers the medical history of 500 patients. All patients are diagnosed with diabetes and visit the hospital in regular intervals for a check-up. During these consultations, they report on their own assessment of their condition (e.g. the self-monitoring of the glucose level) and undergo a set of routine procedures (e.g. the screening for microalbuminuria). The EHR serves to summarize the outcome of these procedures, to give recommendations for further therapy (e.g. a change in insulin dose) or suggest additional interventions (e.g. the transplantation of beta-cells).

A selection of EHRs was manually annotated by five students with a background in biomedical sciences. During the annotation stage, they annotated the complete medical histories of individual patients with clinical codes. They were instructed to identify all medical terms, including non-standard term variants, and link them to the corresponding concept identifier from SNOMED CT (SNOMED International, 2018a). At the end of this stage, the annotation of 179 cases had been completed; 300.693 medical entities were identified, relating to 16,151 unique terms and 7,945 concepts. During the validation stage, three annotators verified the term-concept associations. They were presented with a list of unique pairs of terms and concept codes and asked to confirm the correctness of the assigned code, and rate the domain pertinence of the concept (i.e. whether it is specific to the domain of endocrinology). To estimate the consistency of their judgements, a part of the term-concept pairs was validated by all annotators; the interrater agreement, as calculated by Fleiss' kappa, was substantial (κ=0.62). After filtering out those terms that had been judged as incorrectly annotated during the second stage, we retained 7,687 concepts, corresponding to 15,025 unique terms and 274,082 entities for further analysis.

3.2 Phrase Types included for Analysis

For our study of MWEs, we include nominal phrases with external modifiers as well as compounds. In Dutch, external modifiers can precede or follow the head noun, whereby the grammatical form of the modifier constrains its relative position (Broekhuis, Keizer & den Dikken, 2012). Pre-nominal modifiers can be adjectives (e.g. *lichte hypertensie* 'light hypertension') or participles (e.g. *IgE-gemedieerde allergie* 'IgE(Immunoglobulin E)-mediated allergy'); post-nominal modifiers can be prepositional phrases (e.g. *eczeem aan de handen* 'eczema at the hands'), relative clauses (e.g. *hypertensie die hier wordt opgevolgd* 'hypertension which is followed up here'), participles (e.g. *pijn uitstralend naar schouders* 'pain radiating to shoulders') or adverbials (e.g. *hypoglycemie postprandiaal* 'postprandial hypoglycemia'). Dutch compounds are right-headed. While combinations with other grammatical types

are possible (e.g. *rechterbovenbeen* 'right upper leg'), the juxtaposition of two nouns (e.g. *nierfalen* 'kidney failure') is most productive (Booij, 2007). Many medical compounds contain neoclassical elements, either in combination with each other (e.g. *pancreastransplantatie* 'pancreas transplantation'), or with Dutch lexemes (e.g. *corfalen* 'heart failure'). However, in many neoclassical compounds, the left element (i.e. the non-head) cannot be used as an independent word (e.g. *hypertensie* 'hypertension'). Depending on whether such elements occur in their full or abbreviated form, they have been defined as *confixes* (including *pseudoprefixes* and *pseudosuffixes*), or *splinters* in the literature (Meesters, 2004; Džuganová, 2013). In this study, we do not consider confixes and splinters as separate lexical elements. Consequently, a term such as *hypertensie* will be treated as a single noun, rather than a compound.

3.3 Retrieval of MWEs for Annotation

We focus on MWEs relating to concepts from 2 semantic categories, namely *procedures* and *findings*. To select a set of concepts for each category, we first rank all concepts referring to either *procedures* or *findings* by their absolute frequencies in the annotated part of the corpus. We retrieve the associated terms for the top nine concepts per category. After reviewing the variants linked to one concept, we manually compile a list of lexical stems to retrieve all occurrences of the associated terms. The generation of these stems is based on practical considerations, rather than linguistic criteria. For instance, to retrieve occurrences of the term *hypertensie* 'hypertension', we use the clipped form *hypert* rather than the lexical stem of the headword (*tensie*). The reason is that, firstly, the stem *tensie* would produce a high number of false positives, i.e. matches where the term is used in a different context than that of the target concept (e.g. *tensies thuis* 'tensions at home', which refers to the measurement of the blood pressure in the home setting); secondly, this stem would miss those instances where the clipped variant is used verbatim (e.g. *lichte hypert op rpl*, which is short for *lichte hypertensie op raadpleging* 'light hypertension during consultation').

Then, we match these stems against the entire corpus. For each match, we extract the term along with the three adjacent tokens in the left and right context. Altogether, we identify 63,559 instances of *procedures* and 59,731 of *findings*. Tables 1 and 2 provide examples of the target concepts, terms and lexical stems for both semantic categories.

Concept identifier in SNOMED CT	**Preferred term in SNOMED CT**	**Dutch term variants**	**Lexical stems**
SCTID 73211009	diabetes mellitus	*diabetes mellitus; dm; suikerziekte*	diabet; dm; suikerziek
SCTID 38341003	hypertensive disorder, systemic arterial	*hypertensie; bloedhoogdruk*	hypert; bloedhoogdr
SCTID 414916001	obesity	*obesitas, adipositas*	obes, adipo

Table 1. Concepts, terms and lexical stems used for the identification of MWEs relating to *findings*.

Concept identifier in SNOMED CT	**Preferred term in SNOMED CT**	**Dutch term variants**	**Lexical stems**
SCTID 16310003	diagnostic ultrasonography	*echografie; sonografie*	echo; sono
SCTID 3324009	laser beam photocoagulation	*fotocoagulatie; lasertherapie*	fotoco; laser
SCTID 77465005	transplantation	*transplantatie; tx*	transplant; tx

Table 2. Concepts, terms and lexical stems used for the identification of MWEs relating to *procedures*.

Next, we review the matches to identify all noun phrases where one of the target terms is the syntactic head of the phrase, and where the context conveys medically relevant information. After filtering out those instances that do not fulfill these grammatical and semantic criteria, we retain 11,354 expressions of *procedures*, and 13,537 expressions of *findings*. Finally, we sort the expressions by the relative position of the modifier (i.e. left- vs. right-branching). Among the left-branching expressions, we further distinguish between compound nouns and noun phrases with an external modifier (e.g. a pre-modifying adjective). All compound nouns are split into their constituents for further analysis. This leads to a 3-

way-distinction between phrase types. Table 3 gives an overview of the different phrase types by semantic category.

	Compounds	**Phrases with external modifier in the left context**	**Phrases with external modifier in the right context**
Findings	*ochtend hypo* 'matinal hypoglycemia'	*symptomatische hypoglycemie* 'symptomatic hypoglycemia'	*hypo met convulsies* 'hypoglycemia with convulsions'
Procedures	*YAG laser* 'Nd:YAG laser'	*panretinale laser* 'panretinal laser therapy'	*laser od* 'laser therapy of oculus dexter (right eye)'

Table 3. Overview of the different phrase types by semantic category of the headword.

3.4 Annotation of MWEs with PoS Values and Semantic Features

For the annotation of the MWEs, we use WebAnno, a web-based tool for linguistic annotations (de Castilho et al., 2016). We create two layers of annotation (grammatical and semantic), and define a custom tagset at each level.

Grammatical level: For the tagging of PoS values, we use the tagset introduced in the Penn Guidelines for the annotation of biomedical text (Warner et al., 2012). Compared to the original Penn tagset (Santorini, 1990), this extended version contains 4 additional labels, AFX (unbound affix), GW (mistranscription), HYPH (unbound hyphen) and XX (uninterpretable material). These additional tags have been introduced to handle the particular linguistic features of clinical writing, such as non-canonical spelling variants, ad-hoc abbreviations and undecipherable forms. Similar to Fan et al. (2013), we follow the tagging conventions described in these guidelines, but adjust them to the properties of our data: In the original guidelines, the distinction between common nouns and proper nouns is primarily based on the capitalization of the full form. However, we find that this is not a viable strategy in our case, as orthographic conventions are not followed strictly in our data. Instead, we reserve the tags for proper nouns (i.e. *NNP* for singular, and *NNPS* for plural forms) to eponyms (e.g. *hashimoto*, which refer to 'Hashimoto's disease') and commercial names (e.g. *NovoRapid*, which is the registered tradename of an insulin product). In addition, for abbreviations and misspellings, we used the same tag as for the canonical full form, if it can be determined; otherwise, we tag them with the label for un-interpretable material (*XX*).

Semantic level: For the annotation of conceptual properties, we create a tagset based on the attributes described in the Editorial Guide of SNOMED CT (SNOMED International, 2018b). This guide defines a set of properties that can be used in conjunction with particular semantic types. For instance, a *finding* can be modified with regard to its *clinical course* (e.g. acute, chronic) or *severity* (e.g. mild, moderate); a *procedure* can be specified by the *device* used (e.g. a catheter), or the *direct substance* (e.g. a pharmacological agent used for injection). Semantically empty tokens, such as function words, are skipped on the semantic level. Table 4 provides examples for different types of attributes combining with *findings* and *procedures* respectively.

Attribute	**Example expressions for *findings***	**Attribute**	**Example expressions for *procedures***
Cause	*diabetische retinopathie* 'diabetic retinopathy'	Component	*glycemie meting* 'measurement of blood glucose level'
Severity	*lichte hypertensie* 'mild hypertension'	Site	*schouder ingreep* 'surgery of the shoulder'
Site	*abdominale obesitas* 'abdominal obesity'	Substance	*injectie insuline* 'injection of insuline'

Table 4. Examples of different attribute types combining with *findings* and *procedures*.

3.5 Analysis of Grammatico-semantic Patterns

With our study, we aim to answer two questions: Firstly, we investigate whether there is a correlation between the semantic category of the headword and the preferred phrase types. This will also give an

indication of the average degree of lexicalization of the different categories. Secondly, we examine whether fixed concept combinations pattern with particular grammatical constructions. We thus analyze the annotated MWEs in two stages.

Correlation between semantic categories, preferred phrase types and the degree of lexicalization: For both *findings* and *procedures*, we calculate the absolute and relative frequencies of the individual phrase types, as well as the average length in tokens. For each phrase type, we group the expressions by their paired tag sequences; all expressions that have identical annotations at both the semantic and PoS layer are thus associated with one *pattern*. To evaluate the degree of lexicalization, we count the unique expressions per pattern. This value thus indicates if the associated patterns serve as productive templates, which allow for paradigmatic changes of the lexical elements, or if they correspond to frozen expressions, which are fully lexicalized in usage.

Patterning of concept combinations with grammatical constructions: For a more fine-grained analysis of the grammatical structure, we focus on the five most frequent concept combinations per category. To identify these, we make an inventory of all possible constellations of semantic constituents (e.g. *procedure* and anatomical *site*; *procedure* and *substance*) and rank them by their absolute frequency. For the top five combinations, we extract all patterns (i.e. paired tag sequences) that instantiate these semantic combinations. For instance, for the semantic combination *procedure* and *site*, we identify the patterns 'procedure/NN, site/NN' and 'site/JJ, procedure/NN', whereby the tags 'NN' and 'JJ' refer nouns and adjectives respectively. For the combination of *procedure* and *substance*, we retrieve the patterns 'substance/NN, procedure/NN' and 'procedure/NN, substance/NN'. For each grammatico-semantic pattern, we calculate the frequency relative to all patterns that express the underlying semantic combination.

4 Results

Distribution of phrase types and degree of lexicalization: The distribution of phrase types varies considerably between the semantic categories. While compounds only make up a minor share of the patterns for *findings*, they are the dominant phrase type among *procedures*. Conversely, left-branching phrases with external modifiers are the preferred type for *findings*, whereas they account for a relatively small portion of the *procedures*. For right-branching phrases though, the proportions are almost equal. Regardless of the phrase type, the average length of the expressions is longer for *findings* than for *procedures*. In general, the right-branching expressions are longer than pre-modified phrases and compounds. Overall, the average number of unique expressions by pattern is higher among the *procedures,* which indicates a lower degree of lexicalization. However, even though the relative frequencies of the individual phrase types vary across semantic categories, we note similar global trends with regard to their productivity: For both *findings* and *procedures*, the left-branching phrases are the most productive phrase type, followed by the right-branching phrases. Compounds, on the other hand, appear less variable, hence more lexicalized, for both semantic categories. The full results are provided in Tables 5 and 6.

Patterning of concept combinations with PoS sequences: With both *findings* and *procedures*, we observe a highly skewed frequency distribution of the concept combinations. For both categories, the top 5 combinations of semantic constituents account for roughly two thirds of all identified MWEs. *Findings* are typically specified with regard to their *cause*, *clinical course*, *severity*, the *anatomical site* or *time*. *Procedures* co-occur mostly with modifiers relating to *components*, *abstract properties*, the *time* or *anatomical site*. Likewise, the individual concept combinations are strongly dominated by single PoS sequences. On average, the most frequent PoS sequence accounts for more than half of all MWEs that express a given combination (67.4% for *findings*, 58.3% for *procedures*). However, there are striking differences with regard to the preferred grammatical structure. Among the *findings*, all top patterns consist of left-branching noun phrases, whereby the headword is pre-modified by an adverb, or by one or more adjectives (e.g *veneuze pulmonale hypertensie* 'venous pulmonary hypertension'). By contrast, among the *procedures*, purely nominal sequences prevail; these take either the form of compounds (e.g. *lipiden meting* 'measurement of lipids'), or reduced prepositional phrases, where the subordinating preposition is left out (e.g. *rx thorax* 'x-ray of the chest'). Among the *procedures*, 3 of the 5 concept combinations contain plural forms, which do not appear in the PoS patterns for *findings*. Tables 7 and 8

list the most frequent concept combinations by semantic category, along with the dominant PoS sequence and example expressions.

	Compounds	**Phrases with external modifier in the left context**	**Phrases with external modifier in the right context**
Absolute frequency	187	11,001	2,349
Relative frequency	0.01	0.81	0.17
Average length in tokens	2.84	3.03	3.26
Number of unique patterns	19	383	210
Average number of expressions per pattern	1.33	3.63	2.83

Table 5. Distribution and structure of phrase types among MWEs relating to *findings*.

	Compounds	**Phrases with external modifier in the left context**	**Phrases with external modifier in the right context**
Absolute frequency	7,835	1,482	2,037
Relative frequency	0.69	0.13	0.18
Average length in tokens	2.46	2.46	2.84
Number of unique patterns	392	147	141
Average number of expressions per pattern	2.57	3.69	3.38

Table 6. Distribution and structure of phrase types among MWEs relating to *procedures*.

Concept combination	**Dominant PoS sequence**	**Example expression**	**Relative frequency of the PoS sequence**
finding, cause	JJ, NN	*alimentaire obesitas* 'alimentary obesity'	0.90
finding, clinical course	RB, NN	*vaak hypoglycemie* 'frequently hypoglycemia'	0.35
finding, severity	JJ, NN	*morbiede obesitas* 'morbid obesity'	0.74
finding, site	JJ, JJ, NN	*veneuze pulmonale hypertensie* 'venous pulmonary hypertension'	0.54
finding, time	JJ, NN	*matinale hypo* 'matinal hypoglycemia'	0.83

Table 7. Most frequent concept combinations and PoS patterns relating to *findings*.

In sum, both levels of analysis provide evidence for the interplay of conceptual properties with grammatical structure: The majority of MWEs referring to *findings* consists of left-branching phrases with an external modifier; the grammatical structure of the most frequent concept combinations is nearly identical. By contrast, the *procedures* show a clear preference for nominal constructions, in particular compounds and reduced prepositional phrases. In general, though, MWEs of this category are more variable, which manifests itself in both lexical and morpho-syntactic alternations.

Concept combination	Dominant PoS sequence	Example expression	Relative frequency of the PoS sequence
procedure, component	NNS, NN	*lipiden meting* 'measurement of lipids'	0.44
procedure, component, abstract property	JJ, NNS, NN	*gunstig lipidenprofiel* 'acceptable lipid profile'	0.71
procedure, component, time	NN, NN, NNS	*glycemie dag profielen* 'glycemic day profiles'	0.72
procedure, time	NN, NN	*jaar bilan* 'yearly balance'	0.59
procedure, site	NN, NN	*rx thorax* 'x-ray of the chest'	0.46

Table 8. Most frequent concept combinations and PoS patterns relating to *procedures*.

5 Discussion

Our results illustrate the interdependency of form and meaning in the expression of complex medical concepts. This finding corroborates the observation that "special and specialized information is entrenched in linguistic structures" (Schulze and Römer, 2008).

Firstly, the semantic type of the headword co-determines preferences for particular phrase types. In particular, with left-branching MWEs headed by *procedures*, nominal compounds are most frequent; with *findings*, phrases pre-modified by adjectives prevail. This tendency can be partly attributed to the predominance of a small number of fixed concept combinations. For some types of concepts, the corresponding terms are morphologically inflexible, which enforces the use of particular phrase types. For instance, *procedures* can combine with terms relating to substances, such as *injectie insuline* 'injection of insulin'. For lack of a derived adjective, the use of a nominal construction is – at least in Dutch – obligatory. On the other hand, *findings* are often specified with regard to their clinical course or severity (*morbiede obesitas* 'morbid obesity'), where no semantically equivalent noun form is available. However, this tendency prevails in cases where both grammatical types – nouns and derived adjectives – are available; thus, it cannot be explained by the lack of particular word forms alone. Instead, it seems that grammatical structures are instrumental to convey certain semantic relations. In MWEs relating to *procedures*, the headwords are typically combined with terms referring to concrete entities, which are the direct object of the act itself; this core relation seems to be strongly linked to nominal constructions. On the other hand, *findings* are characterized with regard to inherent properties as they manifest themselves to the observer; such qualities clearly pattern with attributive adjectives. The association between grammatical patterns and semantic relations also accounts for the asymmetrical distribution of reduced phrase types. For instance, the elision of function words is acceptable in some semantic constellations, but not in others. Given the default relation between two nominal constituents ('A is the object of B'), the omission of a preposition is acceptable in phrases combining a *procedure* and an anatomical site, component or substance (e.g. *echografie halsvaten* 'echography of the neck vessels'). By contrast, such an underspecified construction cannot be used for other types of relations. For example, to specify the etiological relation between a cause and a *finding*, the full phrase must be used (e.g. *nefropathie tgv diabetes*, which is short for *nefropathie ten gevolge van diabetes* 'nephropathy resulting from diabetes', but not **nefropathie diabetes* 'nephropathy diabetes').

Secondly, the semantic type influences the variability of the MWEs, both at the lexical, and at the grammatico-syntactic level. Overall, MWEs referring to *findings* are strongly lexicalized. This may be partly explained by the fact that, for the most frequent concept combinations, the set of potential modifiers is rather confined. For instance, for most *findings*, there is only a small number of medically attested causes; this limits the number of combining concepts, and consequently, that of individual expressions. However, even in combination with modifiers of time or severity, which do not underlie such rigid conceptual restrictions, the strong dominance of individual PoS patterns prevails. By contrast, MWEs relating to *procedures* are more flexible; they serve as productive templates that allow for the paradigmatic insertion of different concepts. This increases the potential for variation in the syntactic and grammatical form.

6 Conclusion

MWEs are dense units of information, which enable the concise expression of complex concepts. They play a pivotal role in specialized discourses, as they allow speakers to interact in a precise, yet economical way. Their communicative power resides both in the use of domain-specific terminology, and that of particular constructions, which support the nuanced encoding of meanings and relations. Therefore, the automatic processing of specialized texts crucially depends on their correct interpretation.

The detailed study of specialized corpora is essential to identify such constructions. In this paper, we have presented an analysis of MWEs in clinical usage. Using a structured terminology as a starting point, we have exploited the relations defined in this taxonomy to design an annotation scheme, which allowed us to capture regularities at both the grammatical and the semantic level. While our analysis was confined to a narrow selection of medical concepts, the approach could easily be expanded to a wider range of concepts, or transferred to other domains. Such analyses would lead to valuable insights about the structure of specialized MWEs, which could inform the design of more advanced applications for semantic reasoning.

Acknowledgements

This work was supported by Internal Funds KU Leuven. We are grateful to all anonymous reviewers for their detailed comments and suggestions. We would also like to thank Kristina Geeraert for her valuable input.

References

Afzal, Z., Pons, E., Kang, N., Sturkenboom, M., Schuemie, M. J., & Kors, J. A. (2014). ContextD: An Algorithm to Identify Contextual Properties of Medical Terms in a Dutch Clinical Corpus. BMC Bioinformatics, 15. http://doi.org/10.1186/s12859-014-0373-3

Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W. F. I., Warner, C., Hwang, J. D., … Savova, G. K. (2013). Towards Comprehensive Syntactic and Semantic Annotations of the Clinical Narrative. J Am Med Inform Assoc, 20, 922–930. http://doi.org/10.1136/amiajnl-2012-001317

Baldwin, T., & Kim, S. N. (2010). Multiword Expressions. In N. Indurkhya & F. J. Damerau (Eds.), Handbook of Natural Language Processing (pp. 267–292). Boca Raton: CRC.

Booij, G. (2007). The Morphology of Dutch. Oxford: Oxford University Press.

Bowker, L., & Hawkins, S. (2006). Variation in the Organization of Medical Terms: Exploring some Motivations for Term Choice. Terminology, 12(2006), 79–110. http://doi.org/10.1075/term.12.1.05bow

Broekhuis, H., Keizer, E., & den Dikken, M. (2012). Syntax of Dutch: Nouns and Noun Phrases. Amsterdam: Amsterdam University Press.

Cabezas-García, M., & San Martín, A. (2017). Semantic Annotation to Characterize Contextual Variation in Terminological Noun Compounds: A Pilot Study. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) (pp. 108–113). Valencia: Association for Computational Linguistics.

Choueka, Y. (1988). Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In C. Fluhr & D. Walker (Eds.), Proceedings of the 2nd international conference on computer-assisted information retrieval (Recherche d'Information et ses Applications – RIA) (pp. 609–624).

CLEF eHealth (2018). Datasets. Retrieved May 25, 2018, from https://sites.google.com/site/clefehealth/datasets

Costumero, R., Lopez, F., Gonzalo-Martín, C., Millan, M., & Menasalvas, E. (2014). An Approach to Detect Negation on Medical Documents in Spanish. In International Conference on Brain Informatics and Health (pp. 366–375). Cham: Springer.

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Workshop at the 32nd Annual Meeting of the Association for Computational Linguistics (pp. 29–36). Stroudsburg: Association for Computational Linguistics.

de Castilho, R. E., Mujdricza-Maydt, E., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Proceedings of the LT4DH workshop at COLING 2016 (pp. 76–84). Osaka.

De Hertog, D., & Heylen, K. (2012). The Prevalence of Multiword Term Candidates in a Legal Corpus. In G. Aguado de Cea (Ed.), Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE2012): New Frontiers in the Constructive Symbiosis of Terminology and Knowledge Engineering (pp. 283–290). Madrid: Universidad Politecnica de Madrid.

Deléger, L., Campillos, L., Ligozat, A.-L., & Névéol, A. (2017). Design of an Extensive Information Representation Scheme for Clinical Narratives. J Biomed Inform, 8(37), 1–18. http://doi.org/10.1186/s13326-017-0135-z

Džuganová, B. (2013). English Medical Terminology – Different Ways of Forming Medical Terms. European Journal of Bioethics, 4(7), 55–69.

Faber, P., & León-Araùz, P. (2016). Specialized Knowledge Representation and the Parameterization of Context. Frontiers in Psychology, 7. http://doi.org/10.3389/fpsyg.2016.00196

Faber, P., Tercedor, M., Sánchez, S., López, C. I., León, P., Arauz, A., … Martínez, S. M. (2010). A Cognitive Linguistics View of Terminology and Specialized Language. New York: De Gruyter Mouton.

Fan, J., Yang, E. W., Jiang, M., Prasad, R., Loomis, R. M., Zisook, D. S., … Huang, Y. (2013). Syntactic Parsing of Clinical Text: Guideline and Corpus Development with Handling Ill-Formed Sentences, 20, 1168–1177. http://doi.org/10.1136/amiajnl-2013-001810

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. Language, 64(3), 501–538. http://doi.org/10.2307/414531

Goldberg, A. E. (1995). Constructions. A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.

Goldberg, A. E. (2003). Constructions: A New Theoretical Approach to Language. TRENDS in Cognitive Science, 7(5), 219–224. http://doi.org/10.1016/S1364-6613(03)00080-9

Haverinen, K., Ginter, F., Viljanen, T., Laippala, V., & Salakoski, T. (2010). Dependency-based PropBanking of Clinical Finnish. In Proceedings of the Fourth Linguistic Annotation Workshop (pp. 137–141). Uppsala: Association for Computational Linguistics.

i2b2 National Center for Biomedical Computing (2018). NLP Research Data Sets. Retrieved May 25, 2018, from https://www.i2b2.org/NLP/DataSets/Main.php

Jackendoff, R. (1997). Twistin' the Night Away. Language, 73, 534–559.

Laippala, V., Viljanen, T., Airola, A., Kanerva, J., Salanterä, S., Salakoski, T., & Ginter, F. (2014). Statistical Parsing of Varieties of Clinical Finnish. Artif Intell Med, 61(3), 131–136.

Laso, N. J., & John, S. (2013). A Corpus-based Analysis of the Collocational Patterning of Adjectives with Abstract Nouns in Medical English. In I. Verdaguer, N. J. Laso, & D. Salazar (Eds.), Biomedical English: A Corpus-Based Approach (pp. 55–72). Amsterdam: Benjamins.

Léon, I. K., & Divasson, L. (2006). Nominal Domain in the Biomedical Research Paper: A Grammatico-rhetorical Study of Postmodification. In M. Gotti & F. Salager-Meyer (Eds.), Advances in Medical Discourse Analysis: Oral and Written Contexts (pp. 289–310). Bern: Lang.

Levi, J. (1978). The Syntax and Semantics of Complex Nominals. New York: Academic Press.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2016). Biomedical Term Extraction: Overview and a New Methodology. Information Retrieval Journal, 19(1), 59–99. http://doi.org/10.1007/s10791-015-9262-2

Marciniak, M., & Mykowiecka, A. (2011). Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish. In Proceedings of BioNLP 2011 Workshop (pp. 92–100). Portland: Association for Computational Linguistics.

Meesters, G. (2004). Marginale Morfologie in het Nederlands: Paradigmatische Samenstellinge, Neoklassieke Composita en Splintercomposita. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical Natural Language Processing in Languages Other than English: Opportunities and Challenges, 9(12).

Oleynik, M., Nohama, P., Cancian, P. S., & Schulz, S. (2010). Performance Analysis of a POS Tagger Applied to Discharge Summaries in Portuguese. Stud Health Technol Inform, 160, 959–963.

Oronoz, M., Gojenola, K., Pérez, A., & de Ilarraza, Arantza Díaz Casillas, A. (2015). On the Creation of a Clinical Gold Standard Corpus in Spanish: Mining Adverse Drug Reactions. J Biomed Inform, 56, 318–332.

Pakhomov, S. V, Coden, A., & Chute, C. G. (2006). Developing a Corpus of Clinical Notes Manually Annotated for Part-of-Speech, 75, 418–429. http://doi.org/10.1016/j.ijmedinf.2005.08.006

Ramisch, C. (2015). Multiword Expression Acquisition. A Generic and Open Framework. Cham: Springer International Publishing.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & Setzer, A. (2009). Building a Semantically Annotated Corpus of Clinical Texts. Journal of Biomedical Informatics, 42(5), 950–966.

Santorini, B. (1990). Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd printing). Retrieved from http://www.cis.upenn.edu/~bies/manuals/tagguide.pdf

Savkov, A., Carroll, J., Koeling, R., & Cassell, J. (2016). Annotating Patient Clinical Records with Syntactic Chunks and Named Entities: The Harvey Corpus. Language Resources and Evaluation, 50(3), 523–548. http://doi.org/10.1007/s10579-015-9330-7

Scheurwegs, E., Luyckx, K., Luyten, L., Goethals, B., & Daelemans, W. (2017). Assigning Clinical Codes with Data-driven Concept Representation on Dutch Clinical Free Text. Journal of Biomedical Informatics, 69, 118–127. http://doi.org/10.1016/j.jbi.2017.04.007

Schulze, R., & Römer, U. (2008). Introduction. Patterns, Meaningful Units and Specialized Discourses. International Journal of Corpus Linguistics, 13(3), 265–270. http://doi.org/10.1075/ijcl.13.3.01sch

Smadja, F. (1993). Retrieving Collocations from Text: Xtract. Comput Linguist, 19(1), 143–177.

SNOMED CT. (2018a). SNOMED CT. Retrieved May 13, 2018, from https://www.snomed.org/snomed-ct

SNOMED CT (2018b). SNOMED CT Editorial Guide. Retrieved May 14, 2018, from https://confluence.ihtsdotools.org/display/DOCEG/SNOMED+CT+Editorial+Guide

Styler, W. F. I., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., …Pustejovsky, J. (2014). Temporal Annotation in the Clinical Domain. Trans Assoc Comput Linguist, 2, 143–154.

Sun, W., Rumshisky, A., & Uzuner, Ö. (2013). Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge. J Am Med Inform Assoc, 20(5), 806–813.

ten Hacken, P. (2015). Naming Devices in Midddle-Ear Surgery: A Morphological Analysis. In P. ten Hacken & R. Panocová (Eds.), Word Formation and Transparency in Medical English (pp. 55–72). Newcastle upon Tyne: Cambridge Scholars Publishing.

Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting Medication Information from Clinical Text. Journal of the American Medical Informatics Association, 17(5), 514–518.

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. J Am Med Inform Assoc, 18(5), 552–556.

Warner, C., Lanfranchi, A., O'Gorman, T., Howard, A., Gould, K., & Regan, M. (2012). Bracketing Biomedical Text: An Addendum to Penn Treebank II Guidelines. Retrieved May 14, 2018, from https://clear.colorado.edu/compsem/documents/treebank_guidelines.pdf

Yadav, P., Jezek, E., Bouillon, P., Callahan, T. J., Bada, M., Hunter, L. E., & Cohen, K. B. (2017). Semantic Relations in Compound Nouns: Perspectives from Inter-Annotator Agreement. In A. V Gundlapalli, M.-C. Jaulent, & D. Zhao (Eds.), MEDINFO 2017: Precision Healthcare through Informatics (pp. 644–648). Hangzou: International Medical Informatics Association and IOS Press.