

MAŁGORZATA MARCINIAK

DOMAIN CORPORA
AS A SOURCE OF INFORMATION



INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES

MONOGRAPH SERIES
INFORMATION TECHNOLOGIES: RESEARCH
AND THEIR INTERDISCIPLINARY APPLICATIONS

4

MAŁGORZATA MARCINIAK

**DOMAIN CORPORA
AS A SOURCE OF INFORMATION**



INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES

Warsaw, 2015

Publication issued as a part of the project:
'Information technologies: research and their interdisciplinary applications',
Objective 4.1 of Human Capital Operational Programme.
Agreement number UDA-POKL.04.01.01-00-051/10-01.

Publication is co-financed by European Union from resources of European Social Fund.

Project leader: Institute of Computer Science, Polish Academy of Sciences

Project partners: System Research Institute, Polish Academy of Sciences, Nałęcz
Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences

Editors-in-chief: Olgierd Hryniewicz
Jan Mielniczuk
Wojciech Penczek
Jacek Waniewski

Reviewer: Michał Marcińczuk

Małgorzata Marciniak
Institute of Computer Science, Polish Academy of Sciences
Malgorzata.Marciniak@ipipan.waw.pl
<http://zil.ipipan.waw.pl/MalgorzataMarciniak>

Publication is distributed free of charge

©Copyright by Małgorzata Marciniak

©Copyright by Institute of Computer Science, Polish Academy of Sciences, 2015

ISBN 978-83-63159-12-2
e-ISBN 978-83-63159-13-9

Layout: Piotr Rychlik
Cover design: Waldemar Słonina

Contents

1	Introduction	7
2	Information extraction	9
2.1	Rule based extraction tools	12
2.1.1	Spejd	13
2.1.2	SProUT	14
2.2	Evaluation Measures	16
2.2.1	Evaluation of annotators agreement	16
2.2.2	Evaluation of IE results	16
3	Text corpora	19
3.1	General corpora	20
3.2	Domain corpora	21
4	Medical data anonymisation/de-identification	23
4.1	Regulations and recommendations	23
4.2	Methods	26
5	Construction of medical corpora in Polish	29
5.1	Anonymisation	29
5.2	Conversion into plain text	32
5.3	Document structure	32
5.4	Linguistic characteristics of data	33
5.5	Tokenisation	34
5.6	Morphological analyses	38
5.6.1	Tags	39
5.6.2	Global correction rules	41
5.6.3	Evaluation	43
5.7	Extended tokenisation	46
5.7.1	Complex tokens	46
5.7.2	Method	47
5.8	Semantic Annotation	49
5.9	Corpus structure	50
6	Information extraction from hospital documents	53
6.1	Rule based information extraction	54
6.1.1	Domain description	54
6.1.2	Information representation	55

6.1.3	Domain dictionary	58
6.1.4	Grammar rules	60
6.1.5	Evaluation	66
6.2	Database	69
6.3	Automatic semantic annotation of corpora	71
6.3.1	Method	71
6.3.2	Guidelines for manual verification	74
6.3.3	Evaluation	75
6.4	Machine learning experiment	77
6.4.1	Data	78
6.4.2	Label assignment with CRF model	80
7	Terminology extraction methods	83
7.1	Domain terminology as a set of terms	84
7.2	Terminology extraction process	86
7.3	Variants recognition	89
7.4	Examples of terminology extraction methods	92
7.4.1	C/NC method	92
7.4.2	A statistical term extractor	93
7.4.3	Contrastive measure	95
7.4.4	Method using domain-specificity and term cohesion	96
7.4.5	Term extraction from sparse data	97
8	Terminology extraction from hospital documents	99
8.1	Term phrases	100
8.1.1	Phrases description	100
8.1.2	Shallow grammar	102
8.2	Simplified base forms	104
8.3	Candidate terms ranking	106
8.4	Results	109
8.4.1	Statistics	109
8.4.2	C/NC phrase reordering	112
8.4.3	Manual annotation	112
8.4.4	Manual evaluation	114
8.4.5	Comparison with Polish MeSH	116
8.5	NPMI driven recognition of nested terms	118
8.5.1	Motivations	119
8.5.2	Algorithm	119
8.5.3	Examples	121
8.5.4	Statistics of phrases	126
8.5.5	Evaluation	127
8.5.6	Three-word phrases	128
8.6	Comparison with a general corpus	129
8.7	Converting simplified forms into terminology resources	133
8.8	Towards domain ontology construction	136

9 Summary	139
Appendix A Polish tagset	141
Appendix B Example of text annotation	143
Appendix C Shallow grammar for phrase recognition	147
Appendix D Top 100 medical phrases	151
References	155
Index	169

Preface

The goal of this book is to show what kind of information can be obtained from domain corpora and how to obtain it. The volume consists of three parts devoted to three related issues: domain corpora creation and annotation, relation extraction, and terminology extraction. The problems dealt with are focused on the processing of Polish texts. As an exemplary domain, I have chosen medical data. Taking this decision I was guided by the number of interesting issues to be solved in processing medical documents, as they are significantly different from general texts. The ideas and results presented in this book have previously been published in conference and journal papers.

The book is meant for readers with a computer background and who are interested in natural language processing. It is written with a presumption that readers' knowledge of Polish grammar is basic, but I assume that they are familiar with parts of speech and grammar cases.

The book is published within the project "Information technologies: research and their interdisciplinary applications" that imposes English as the language of publications. All issues are illustrated with examples in Polish for which an English translation is given, and, where it is helpful, the word to word equivalent is added. I hope it makes the problems comprehensible for non-Polish speakers too.

I would like to express my gratitude to my colleagues from the team as all the results described in the book were obtained in collaboration with Agnieszka Mykowiecka, and part of the programming work was performed by Piotr Rychlik. I would like to thank Jakub Piskorski for tips on how to use the SProUT system and Anna Kupś who worked with us on the first experiments on information extraction from mammogram reports. Last but not least, I would like to express my gratitude to the late Professor Leonard Bolc. He first inspired us to address the problems of information extraction. He always encouraged us to write books, because the dissemination of scientific results was his mission.

Finally, I would like to thank the first readers of this book, Piotr Rychlik and Agnieszka Mykowiecka, who suggested changes to make the book easier to read, and to Terry Deal who corrected the English. They helped me to improve and smooth the text. All deficiencies that remain in the book are entirely my responsibility.

Introduction

Domain corpora are collections of documents related to a selected domain and representative of that domain. Such data is useful for exploring knowledge about selected facts described in texts, as well as for learning vocabulary and terminology specific to the domain, or even for creating the domain description. The subfield of natural language processing, devoted, among other things, to these problems, is called information extraction.

Information extraction (IE) is the automatic identification of predefined types of information like entities, relations, or events in free texts; the output is limited to the target information. The similarly named subdomain of natural language processing — Information Retrieval (IR) — is focused on finding documents containing given information and complete documents are the result of the search.

IE deals with several subtasks listed below. Named Entity Recognition (NER) is the task of recognition and classification of named entities as persons, organisations, or geographic locations as well as time descriptions or numbers (money or measure values). Named entity recognition might be a part of the relation extraction task which aims at looking for relations between objects, for example to look for people being employed in institutions, or to establish where institutions have their head office. When we extract more complex information we are talking about event extraction. For example, if we extract information concerning transactions on a stock exchange, we might want to establish their details, such as the investor, subject of transaction, date, amount of shares, location, etc. IE also includes terminology extraction which is aimed at recognition vocabulary and phrases typical to a domain. Finally, coreference resolution is also considered as an IE task; the problem concerns linking text entities describing the same objects in the real world. In the book, we look at some IE problems in the context of domain corpora and we concentrate on relation/event extraction and terminology extraction from domain text collections.

At the beginning of the book, a brief introduction to corpora, their types, and several examples are given. Then, we focus on the processing of medical texts, in particular, hospital records, so we direct our attention to restrictions on collecting hospital records and aspects of protecting personal data contained in such documents.

In the next chapter, we describe how to create medical corpora in Polish. Choosing the medical domain we were guided by the number of interesting issues to be solved in processing such texts. This is interesting, as each step of linguistic analysis contains problems that are usually not present in the case of other data. Medical documents are often noisy; the vocabulary differs significantly from the

general Polish; texts contain typographic errors, spelling variants of words, and abbreviations (some of them are created ad hoc). The resources and tools for processing Polish are not sufficiently developed, especially those dedicated to domain text processing. The scientific community engaged in natural language processing in Poland aims to develop new resources and tools that would make easier to solve some problems described in the book.

The next part concerns extraction of complex information with a detailed discussion of a rule based approach to information extraction from hospital records. Nowadays, creating rules manually is a classical approach to the task, but it is commonly used, efficient, easy to maintain and modify, so worth presenting in our opinion. In this chapter, we also describe how to use the extracted results: to create a database, to prepare data for a machine learning system, and to prepare a semantic annotation layer of a corpus.

The next two chapters are devoted to terminology extraction. First, we describe the problem and present a general approach to this task. Then, we discuss how to use one of the most widely applied methods — the C/NC method (Frantzi *et al.*, 2000) — to Polish data. Then, we put forward our modification that prevents the creation and promotion of truncated nested phrases being considered as terms. Finally, we make a few comments on the problem of term clustering in order to obtain semantically coherent groups, which is the first step towards automatic extraction of a domain ontology.

Information extraction

Information extraction, as a separate subdomain of NLP, started to develop in the 1980s. The series of Message Understanding Conferences (MUC), held in 1987–1997, was designed to promote research in information extraction, see (Grishman and Sundheim, 1996). The conferences were competition-based where many research teams solved the same problem, such as recognition of terrorist activities in Latin America (the topic of MUC-3 and MUC-4). As the results obtained by teams had to be compared, the standards for evaluation were developed.

The first information extraction systems were based on manually created regular expressions and lexicons. Preparation of such extraction grammars requires qualified engineers with linguistic knowledge who know how to create patterns effectively. The extraction process is often organised into a cascade of grammars (e.g. finite-state transducers). The grammar is usually a part of a larger system that provides components typical for language processing like tokenisation, morphological analysis (sometimes together with tagging), division into sentences, and the possibility of adding specialised lexicons called gazetteers; see (Hobbs and Riloff, 2010). Manually prepared systems are effective but their creation is considered time-consuming, and certainly they are domain dependent. In spite of that, information extraction methods based on regular expressions are often used in processing medical data ((Meystre *et al.*, 2008), (Gold *et al.*, 2008)) due to their high precision and effectiveness.

In the next stage of IE development, researchers started to apply statistical methods to information extraction. These techniques require preparing data for training systems; but, when we have such data, we can use ready algorithms based on e.g. Hidden Markov Models (HMMs) (Seymore *et al.*, 1999), (Freitag and McCallum, 2000) or Support Vector Machines (SVM) (Han *et al.*, 2003), (Li *et al.*, 2005). Lafferty *et al.* (2001) proposed using Conditional Random Fields (CRF) to label sequence data, and below we quote a fragment of the abstract of their paper that indicates the advantages of the CRF method:

“We present conditional random fields, a framework for building probabilistic models to segment and label sequence data. Conditional random fields offer several advantages over hidden Markov models and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models. Conditional random fields also avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed

graphical models, which can be biased towards states with few successor states.”

In natural language processing, CRF is applied to various tasks: shallow parsing (Sha and Pereira, 2003); named entity recognition in general (McCallum and Li, 2003) and from the biomedical domain (McDonald and Pereira, 2005); for information extraction focused on form filling (Peng and McCallum, 2004), (Kristjansson *et al.*, 2004). The CRF method is, nowadays, the most popular machine learning approach to IE problems. It is, in particular, the standard approach to the named entity recognition.

Contrary to expectations, machine learning methods don’t significantly reduce the manual work necessary to prepare information extraction systems. The manual work of qualified engineers preparing grammars is substituted by the manual work of annotators, as each application requires its own annotated data.

As building rules manually is the classical approach to the IE task, only 3.5% of conference papers on information extraction in the 10 year period between 2003 and 2012 concerned rule based systems according to Chiticariu *et al.* (2013). They counted that 75% of papers in this period concerned the machine learning approach, while 21% — a hybrid approach. The same authors state “The rule-based approach, although largely ignored in the research community, dominates the commercial market.” This popularity is due to the fact that such systems are easy for developers to understand; hence, they can easily customise a system to new needs and correct errors. This is not the case for machine learning based systems. The Chiticariu *et al.* summary of the advantages and disadvantages of both approaches is given in Table 2.1.

Table 2.1: Pros and Cons of IE systems, (Chiticariu *et al.*, 2013)

Methods	Pros	Cons
Rule based	<ul style="list-style-type: none"> • Declarative • Easy to comprehend • Easy to maintain • Easy to incorporate domain knowledge • Easy to trace and fix the cause of errors 	<ul style="list-style-type: none"> • Heuristic • Requires tedious manual work
ML based	<ul style="list-style-type: none"> • Trainable • Adaptable • Reduce manual effort 	<ul style="list-style-type: none"> • Requires labeled data • Requires retraining for domain adaptation • Requires ML expertise to use or maintain • Opaque

The idea of facilitating construction of IE grammars by the bootstrapping of extracting patterns allows manual work to be reduced. It is not new, as Riloff proposed in (1996b) a method for using domain-independent linguistic rules (general heuristic patterns) to construct extraction patterns automatically. For example, for the terrorism domain, on the basis of annotated noun phrases in texts and the initial pattern: “<subject> passive-verb”, it is possible to create the rule “<victim> was murdered”. Then, the same author, (Riloff, 1996a), proposed extending a pattern based extraction system with new patterns on the basis of a corpus with indicated relevant and irrelevant phrases.

A bootstrapping idea, presented in two papers (Gupta and Manning, (2014b; 2014a)), is implemented in the open domain SPIED (Stanford Pattern-based Information Extraction and Diagnostics, <http://nlp.stanford.edu/software/patternslearning.shtml#About>) tool. The authors apply learning methods to improve patterns used in rule based information extraction. So their system combines advantages of both approaches, rule-based and machine learning, as obtained patterns can be manually corrected.

For large corpora, unsupervised information extraction methods can be applied. They don't need any annotated data or manually created rules. Hasegawa *et al.* (2004) addressed the idea of unsupervised recognition of important relations among named entities. The idea consisted in the following five steps quoted from the paper:

- tagging named entities in text corpora;
- getting co-occurrence pairs of named entities and their context;
- measuring context similarities among pairs of named entities;
- making clusters of pairs of named entities;
- labelling each cluster of pairs of named entities.

Predicate:	Class1
Pattern:	NP1 “such as” NPList2
Constraints:	head(NP1)=plural(label(Class1)) & properNoun(head(each(NPList2)))
Binding:	Class1(head(each(NPList2)))

Fig. 2.1: General rule in KNOWITALL, (Etzioni *et al.*, 2005)

A well-known example of the unsupervised approach to information extraction is the KNOWITALL (Etzioni *et al.*, 2005) system that extracts named entities, and instances of relations (like those recognising the capital of a country) from the Internet. It uses generic, domain-independent rules that are tuned to specific information we are looking for. For example, if the system has to collect names of cities, it tunes the general rule in Figure 2.1 setting the head¹ of the

¹ The head element of a phrase is the word around which the whole phrase is built. It determines its syntactic description.

NP1 noun phrase to ‘City’ and extracts a list of city candidates. Then, the system uses statistics methods (based on pointwise mutual information) for testing the plausibility of these candidates.

Banko *et al.* (2007) introduced the Open Information Extraction (OIE) notion. The idea is to find all possible tuples $t = (e_i, r_{i,j}, e_j)$, where e_i and e_j denote entities and $r_{i,j}$ represents a relationship between them. For example, the entities might be represented by noun phrases that constitute the subject and the object in a sentence, while a relation might be a verb phrase joining them. Then, relations are normalised and assigned probabilities calculated on the basis of the number of distinct sentences from which they were extracted. If our task consists in looking for a piece of information — a query — we try to find a tuple that matches elements of the query, and probabilities help to find the best matching. The process can be more effective if we use synonymic relations. Open information extraction is applied to many tasks such as: learning inference rules from Web text (Schoenmackers *et al.*, 2010), automatic acquisition of commonsense knowledge (Lin *et al.*, 2010), or ontology learning (Poon and Domingos, 2010) and population (Koukourikos *et al.*, 2012).

2.1 Rule based extraction tools

In the book, we focus on methods of creating and exploiting domain corpora, so we don’t describe IE tools in detail but mention some of them together with relevant publications.

The most widely known system is GATE (General Architecture for Text Engineering, <https://gate.ac.uk/>, Cunningham *et al.* (2011)) — the open domain language processing tool which is designed, among other things, to extract specified events, entities or relationships from unrestricted texts. It is “capable of solving almost any text processing problem” as the authors advertise it — its guide consists of almost 600 pages. For English, it provides tools for tokenisation, development and application of gazetteers, splitting sentences, tagging, named entities recognition and coreference identification. Some of these functions are available for several other languages too. Its JAPE (Java Annotation Patterns Engine) grammar consists of pattern matching rules operating on annotated texts and as results modifying these annotations or attaching a new one.

SProUT (Shallow Processing with Unification and Typed Feature Structures) system (<http://sprout.dfki.de/>, Drożdżyński *et al.* (2004)), is a general purpose platform consisting of a set of components for basic linguistic operations: tokenisation, sentence splitting, morphological analysis and domain dictionary analysis on the basis of the defined gazetteer. The SProUT grammar formalism combines finite-state techniques and unification-based formalism that generate Typed Feature Structures (TFS) from text. The system is available for non-commercial use after signing a license with DFKI GmbH, the institution where the system was developed.

ExPRESS (Piskorski, 2008) provides a rule specification language that combines features of JAPE developed in the GATE system with formalism deployed

in SProUT. To make the system easier and more effective the authors resign from the unification operation available in SProUT. ExPRESS is designed in order to process a huge amount of data quickly, so the efficiency of the system is its important feature. The system was developed in the Joint Research Centre of the European Commission and is available for research purposes after signing an agreement with the author.

Another open domain information extraction tool is TEXTMARKER (<http://sourceforge.net/projects/textmarker/>, Kluegl *et al.* (2008)). The system provides a transformation-based approach, where each rule works on existing annotations in order to create new ones. Rules are applied in the order of their listing. The system provides an environment that supports testing rules and their creation. An interesting aspect of the system is “the usage of scoring rules for uncertain and heuristic extraction” (Kluegl *et al.*, 2008) as it is possible to add scoring points to the matched fragments (MARK action), which might then be evaluated by the SCORE condition. This procedure is applied to create new annotations.

Spejd is the open domain system (<http://zil.ipipan.waw.pl/Spejd/>, (Przepiórkowski, 2008), (Buczyński and Przepiórkowski, 2009)). It is a cascade of regular grammars where each single rule constitutes a regular grammar. The primary intended application was morphosyntactic disambiguation and shallow parsing, i.e. phrase recognition.

In our work, we use the two extraction systems mentioned above, namely Spejd and SProUT. The first one we applied to recognise complex tokens like dates and decimal fractions. The second is applied to extract relations from a medical corpus, and creates a semantic annotation of the corpus. The annotation contains selected information about patients’ illnesses. Below we give an example of a rule in these formalisms to explain their syntax necessary to understand the examples given in Section 5.7 and Section 6.1.4.

2.1.1 Spejd

In our works, we used the only version of the tool available then, i.e. the version 0.84 implemented in Java, as the new Spejd (Zaborowski, 2012) was released in 2012 when our work was completed. As the task we wanted to resolve consisted in recognition and annotation of groups of tokens, even the previous Spejd was suited for carrying out the task.

Each rule operates on the results of rules included earlier in the rule file and consists of the following elements:

- **Rule** part indicates a name;
- **Eval** part describes the created element, its type and value, e.g. `word(type, value)`; and operations on elements described by the rule, e.g. `unify(case number gender, 1, 2)`;
- **Mach** part describes elements to which a new element is assigned;
- **Left** and **Right** parts specify the context of the match, which might be empty and then omitted.

An example of a simple Spejd rule from (Buczyński and Przepiórkowski, 2009) is given in Figure 2.2. The rule recognises a group consisting of two tokens described in the **Match** part of the rule. So, the group consists of a preposition (`pos~"prep"`) and a lexeme that has the base form *co* ‘what’ or *kto* ‘who’. In the **Eval** part, the group (**PG**) consisting of both tokens is created if the tokens have interpretations with the same cases — `unify(case,1,2)`. The numbers refer to subsequent tokens described in the **Left**, **Match** and **Right** parts of the rule, so number 1 refers to the preposition while 2 to the *co/kto* element. Spejd assumes that tokens are separated by spaces, but it is possible to indicate lack of a space using the `ns` token.

```
Rule "example-rule"
Left: ;
Match: [pos~"prep"][base~"co|kto"] ;
Right: ;
Eval: unify(case,1,2); group(PG,1,2) ;
```

Fig. 2.2: Rule identifying a prepositional group

Spejd provides several predicates: e.g. `agree` checks agreement; `unify` not only checks agreement but also deletes incompatible interpretations; operations that delete, leave or add interpretations. Moreover, it is possible to create a group of words (`group` operation) or a new syntactic word (`word` operation).

```
Rule "new-example-rule"
Match: A [pos~"prep"] B [base~"co|kto"] ;
Eval: unify(case,A,B); group(PG,A,B) ;
```

Fig. 2.3: Rule identifying a prepositional group in the new Spejd

The new version of Spejd contains many improvements. The authors allow for using references to specification units (numbers of units can still be used) to indicate tokens which take part in operations. The rule in Figure 2.3 uses references instead of numbers. The new version of Spejd allows for specification of a sequence of entities which may occur between any pair of specification units (the new **Between** section). It allows more relaxed rules to be written, for example, it is possible to easily omit punctuation marks between elements. Moreover, new operations are introduced and it is possible to assign a value of an expression to a given variable.

2.1.2 SProUT

The grammar formalism in SProUT consists of rules, which are regular expressions (recognition patterns) over typed feature structures with unification. Output structures are also TFSs.

The domain model used in SProUT is represented by a multi-hierarchy of TFSs (Emele, 1994). Every TFS has a name of a type assigned and a set of features (attribute names and values). A feature's value can be an atomic type, another TFS, or a list of atomic types or TFSs.

Rules refer to three sources of information:

- *token* indicates what kind of characters are in an elementary string. This information is used for recognising, among other things, abbreviations, dates, and numbers.
- *morph* for representing morphological interpretations. In the case of Polish, the Morfeusz analyser (Woliński, 2006) is integrated with SProUT.
- *gazetteer* for a description of entries from a domain dictionary. It can contain all forms of terms important to a domain terminology.

SProUT allows us to use only one source of information for a string interpretation in a rule.

Each rule begins with a rule name. Then, after the ':>' symbol, the rule body coding a regular expression is defined. It describes a sequence of elements which must be identified in the text to trigger the rule. After the '->' symbol, the resulting output structure is given. The rule can contain an alternative ('|'), optional ('?') and repeated ('*') element. Moreover, it is possible to refer to the results of previously defined grammar rules with the `@seek` operator. Text is processed from the beginning to the end and each token can be recognised by only one rule. If several rules recognise the same part of text the result comes from the rule that covers the longest text.

example-rule :>

```

  morph & [POS Prep, INFL [CASE_PREP #c]
    (morph & [STEM "co", INFL [CASE_NOUN #c]] |
     morph & [STEM "kto", INFL [CASE_NOUN #c]])
-> phrase & [TYPE pg].

```

Fig. 2.4: Rule identifying a prepositional group in SProUT

Let us consider an example in Figure 2.4 that contains a rule recognising the same phrase as the *Spejd* rule in Figure 2.2. The rule refers to two subsequent elements, where the second is described as the alternative of two nouns with the specified base forms. Both elements refer to morphological descriptions from the dictionary. The first one is a preposition while the second one is described by the base form (STEM). The unification of cases is realised by the same variable *#c*. As a result, we obtain the TFS of the *phrase* type, for which we define one attribute *TYPE* with the value *pg*.

2.2 Evaluation Measures

Evaluation measures are indispensable for comparing methods and tools applied to the same task. They are also necessary for monitoring the effectiveness of modifications introduced to a method or a tool. Below, we describe some popular measures applied to evaluate results of information extraction tasks, and a measure used to evaluate the quality of test data.

2.2.1 Evaluation of annotators agreement

If we want to evaluate the results of any extraction system, we have to prepare correctly annotated data — a test set which is either prepared manually or manually verified. The results obtained by the program are compared to the test set to evaluate the performance of the system. To obtain a high quality test set, annotation tasks are usually performed by at least two instructed annotators. The final annotation is negotiated between the annotators or another person resolves conflicts. Such a procedure provides a high quality final annotation. If we want to check whether a task is well formulated, we count the κ measure, defined in Equation 2.1, which represents the correlation between the scores of two annotators. It takes into account not only how often the annotators agreed in their rating but, also, how often this agreement takes place by chance.

$$(2.1) \quad \kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

where $Pr(a)$ is the observed percentage of agreement, while $Pr(e)$ is the hypothetical probability of chance agreement.

The observed percentage of agreement is the proportion of ratings where annotators agree in their ratings. So, it is the number of the same ratings to the number of all ratings. The expected percentage of agreement is the proportion of agreements that might be expected “by chance” if the annotators are scoring randomly. For each answer, we check the probability of it being chosen by both annotators and multiply it by the probability of this answer being chosen by chance, and we sum up all these results to obtain the probability of chance agreement.

The κ measure values are from ranges $< -1, +1 >$, where +1 indicates the perfect agreement between the annotators, while -1 indicates the perfect disagreement of them. A κ value of 0.70 is usually considered as reliable, but in (Artstein and Poesio, 2008) the authors conclude that “only values above 0.8 ensured an annotation of reasonable quality”.

2.2.2 Evaluation of IE results

The basic measures used in the evaluation of searching tasks, e.g. information extraction or information retrieval, are *precision*, *recall* and their combination *F-score*. In Equations 2.2–2.4 we define *precision*, *recall* and *accuracy* measures.

These measures are defined on sets of items that are relevant or irrelevant to the task. They are divided into four classes, see Table 2.2. The set of relevant and found/extracted items is called *True Positives*. Incorrectly identified items create the *False Positive* set. The correct but unfound items create the *False Negative* set, and incorrect and unidentified items create the *True Negative* set. In measures defined in Equations 2.2–2.4, we use the cardinality of these sets.²

Table 2.2: Classification of items

	Relevant to the task	Irrelevant to the task
Found	<i>True Positive</i>	<i>False Positive</i>
Not found	<i>False Negative</i>	<i>True Negative</i>

$$(2.2) \quad \textit{precision} = \frac{|True\ Positive|}{|True\ Positive| + |False\ Positive|}$$

$$(2.3) \quad \textit{recall} = \frac{|True\ Positive|}{|True\ Positive| + |False\ Negative|}$$

Precision and *recall* do not refer to the *True Negative* set. They can therefore be used in tasks that extract phrases. For these tasks it would be impossible to determine the cardinality of the set of all possible to extract phrases. The *True Negative* set is taken into account in the *accuracy* measure defined in Equation 2.4.

$$(2.4) \quad \textit{accuracy} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

Precision and *recall*, considered separately, does not show how good the method is. If, for a task, we obtain a result consisting of only one correct item (from a large amount of other correct items) the *precision* is maximal, i.e. 1 (or 100%), but the *recall* is low. If our method identifies all items regardless of whether they are correct or not, the *recall* is maximal but precision is low. These two measures are often combined into one measure. The F_β -score is defined in Equation 2.5 and its special case *F-score* in Equation 2.6.

$$(2.5) \quad F_\beta\text{-score} = \frac{(1 + \beta^2) * \textit{precision} * \textit{recall}}{\beta^2 * \textit{precision} + \textit{recall}}$$

where β is a real positive number.

F_β -score weights higher *precision* for β greater than 1, and *recall* for $\beta \in (0, 1)$ while for $\beta = 1$ *precision* and *recall* are evenly weighted and we obtain

² In Equation 2.4 we use acronyms.

the measure defined in Equation 2.6. Hereinafter, if we refer to the *F-score* we understand this last balanced measure.

$$(2.6) \quad F\text{-score} = \frac{2 * \textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

Text corpora

Linguistic corpora are collections of written text and/or recorded speech which are stored in an electronic form. They are created in order to study various aspects of languages. A well designed corpus should be constructed according to several criteria of text selection to obtain the resource representative for a language, a language genre, or a specific domain. Each corpus should contain real text examples coming from books, newspapers, journals, leaflets, blogs, chats, etc. in a proportion that represents their usage. Corpora that fulfil this intuitive definition are called representative and balanced. The important aspect in collecting data is to check up on copyright issues. Intellectual property rights differ slightly in various countries, but it is usually a big obstacle, at least in making corpora publicly available.

There are different criteria for corpora classification that take into account types of data included in corpora. Corpora may contain texts, speech or both types of data. They may represent data concerning one or several languages. Multilingual corpora might be parallel or comparable — parallel corpora contain translated texts while in comparable corpora, texts relate to the same subject but are independently written in each language, for example, news describing the same event. General corpora contain general texts and allow us to study a language as it is, while specialised corpora contain data from a chosen domain or a language genre (e.g. language of Internet fora) to study the language of a particular domain or a group of people.

In order to prepare a corpus, collected data is converted into plain text and annotated with various types of linguistic information. First, an initial linguistic analysis is performed. Text is segmented into smaller parts like paragraphs, sentences and tokens. For languages with rich inflection, like Polish, morphological analysis is indispensable for further text processing. Taggers, cooperating with morphological analysers, assign descriptions to each token. In the case of Polish words, the description consists of a word lemma, a part of speech, and an entire morphological characterisation. Corpora can be also annotated with a variety of information such as: named entities, coreferences, concepts, relations, phrases.

There are many publications concerning the problem of corpus construction and annotation. The book by Wynne (2005) provides an introduction to the construction of linguistic corpora for beginners. A short overview of problems concerning corpora creation is given in (Xiao, 2010), while 61 papers included in the handbook (Lüdeling and Kytö, 2008) provide an exhaustive survey of almost all aspects of corpus linguistics, its history, applications and correlation with other domains.

Some overviews of corpora are available in print, see (Xiao, 2008) or (Hajnicz, 2011). Several Internet pages include references to corpora and their descriptions, see http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/index2.html or <http://linguistlist.org/sp/GetWRListings.cfm?wrtypid=1>. A few examples of corpora are given in the rest of the chapter.

3.1 General corpora

British National Corpus (BNC) (Burnard, 2007) is a widely known example of a general corpus. It consists of about 100 million words. The corpus was created in 1991–1994, and after that no new data was added but the corpus and its annotation was revisited. The corpus is publicly available as all data (texts and spoken material) has a permit obtained from the owners of the rights to be included in the corpus.¹ The corpus contains British English texts (90%) and transcribed speech (10%) from the last two decades of the 20th century. Representativeness is provided by careful selection of diverse texts according to: domain, time of creation, and medium of publication. The current version of the corpus schema complies with TEI guidelines (Sperberg-McQueen and Burnard, 2008). Texts are annotated with metadata e.g.: authors, titles, time of creation. The structure of texts is represented by headings, paragraphs, lists, sentences, etc. Each word is labelled with its part of speech.

Polish National Corpus (Narodowy Korpus Języka Polskiego, NKJP, <http://nkjp.pl/>, Przepiórkowski *et al.* (2012)), created in 2008–2010, consists of 1.8 billion words. Its balanced subcorpus (modeled on BNC) contains 300 million words. All data can be searched via the Internet. The corpus annotation consists of several layers. Data is segmented into sentences, tokens representing words are annotated with morphosyntactic information (parts of speech and grammatical features). The other levels of annotation contain:

- syntactic words, e.g. analytical forms (*będę szedł* ‘I will go’), multi-segment adverbs (*po ciemku* ‘in the dark’);
- syntactic groups’ e.g. nominal groups (*duży dom* ‘big house’), numerical groups (*dwa domy* ‘two houses’), prepositional groups (*za domem* ‘behind the house’);
- named entities, e.g. proper names of persons, geographical objects and organisation, and temporal expressions.

All annotations were automatically done with tools specially prepared for these tasks. The tools are now publicly available. A 1-million-word balanced subcorpus consisting of short extracts was created from the 300 million balanced corpus and made fully available (short extracts are exempted from copyright restrictions). For this data the annotations were manually corrected.

¹ Examples of letters asking for such permits, the authors of BNC published on <http://www.natcorp.ox.ac.uk/corpus/permletters.html>.

3.2 Domain corpora

Domain corpora are fundamental resources for research concerning creation of domain dictionaries, thesauri, and domain ontologies. They may consist of a set of journal abstracts, journal papers, books, or other domain texts. Domain corpora may be collected on the Internet, from publishers or existing archives, or in other places where text documentation is created, such as hospitals in the case of medical data.

The majority of domain corpora available on the Internet concerns the biomedical domain. They usually consist of the abstracts of biomedical papers or full papers. Biomedical corpora annotated with various types of information and created in order to prepare and test various applications are available from e.g. <http://www.nactem.ac.uk/resources.php>.

GENIA (<http://www.nactem.ac.uk/genia/>) is the best known biomedical corpus. It consists of 1999 Medline² abstracts of articles related to the following MeSH³ terms: *human*, *blood cell*, and *transcription factor*. The corpus contains approximately 500,000 tokens. GENIA is annotated with various levels of linguistic and semantic information. As the data is very different from everyday English (it contains a lot of proper names, abbreviations, chemical and numerical expressions, a lot of strings with hyphens and slashes) its tokenisation is difficult. Teteisi and Tsujii (2006) formulated guidelines for tokenisation and part of speech tagging of the corpus and other biomedical texts. The authors describe how to use the Penn Treebank (PTB) annotation scheme (unofficial standard for English) and what changes are necessary in tokenisation and POS assignment. Two experts manually annotated the data with terminology, they indicated almost 100,000 terms. Several experiments concerning automatic terminology extraction were evaluated on this data, e.g. (Zhang *et al.*, 2008), (Knoth *et al.*, 2009), (Lossio-Ventura *et al.*, 2014). Moreover, the corpus is annotated with syntactic trees (Tateisi and Tsujii, 2006), events (Kim *et al.*, 2008), relations (Ohta *et al.*, 2010) and coreferences (Su *et al.*, 2008).

For Polish, two economic corpora are available via the Internet. They are automatically annotated with morphosyntactic information and manually annotated with a word sense layer, see Kobyliński (2012). Both corpora were collected in order to perform experiments with word sense disambiguation. A set of stock market reports collected from the Internet make the gpwEcono corpus available from <http://zil.ipipan.waw.pl/gpwEcono>, it contains approximately 300,000 tokens. The plWikiEcono corpus <http://zil.ipipan.waw.pl/plWikiEcono> consists of 1219 economic articles (less than 1 million tokens) from Wikipedia.

² Medical Literature Analysis and Retrieval System Online (<http://www.nlm.nih.gov/pubs/factsheets/medline.html>) is a database of bibliographic information of papers concerning biomedical domain and life science.

³ Medical Subject Headings (<http://www.nlm.nih.gov/pubs/factsheets/mesh.html>) is a thesaurus organised in hierarchical structure and containing biomedical domain and life science terminology created in order to index papers.

Medical data anonymisation/de-identification

The book addresses the problems of processing domain corpora, especially those consisting of medical documents. Thus, special attention should be paid to the protection of personal data. This chapter contains an overview of regulations, recommendations and methods concerning the problems.

The collecting of clinical data needs to preserve the patients' right to protect their personal data against inappropriate disclosure. Due to the growing importance of computer technologies in processing medical data (e.g. storing medical documentation in computers, exchanging it on computer networks, processing by various programs), legal rules defining personal data and methods of protecting it have been developed. Documents must be de-identified (or anonymised) before they are made accessible for further processing. There are two aspects of this problem: the first one concerns regulations that should be familiar to those involved in processing such data, while the second one concerns the methods for fulfilling these requirements. Both are described briefly in this section.

4.1 Regulations and recommendations

The growing importance of computer technology in everyday life, including medical care, has resulted in the development of legal standards for protecting personal information. Initially, in the 1990s, regulations on the protection and processing of personal data were developed in many countries. In the European Union (EU), *Directive 95/46/CE of the European Parliament and of the Council of 25 October 1995 on the protection of individuals with regards to the processing of personal data and on the free movement of such data*¹ was published. In Poland, these issues are regulated by the Act of 29 August 1997 on the protection of personal data (*Ustawa o ochronie danych osobowych*, <http://isip.sejm.gov.pl>). In 2014, the EU, in cooperation with the Council of Europe (CoE), published the *Handbook on European data protection law*. It contains an exhaustive review of European regulations concerning data protection (including medical documentation) and contains answers to legal questions that are illustrated with specific cases.

In the USA, the law specifically targeted at health care appeared as *The Healthcare Portability and Accountability Act of 1996* (HIPAA), available from

¹ On the page <http://eur-lex.europa.eu/browse/summaries.html> see the topic: Information society/Data protection, copyright and related rights.

<http://privacyruleandresearch.nih.gov/pdf/>. It regulates the way to protect health information, what kind of information is considered as Protected Health Information (PHI) and how health care professionals and providers should deal with patient information. Moreover, the Department of Health and Human Services (HHS) issued the Privacy Rule in December 2000 to carry out defined standards, see the booklet prepared for researchers *Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule* (<http://privacyruleandresearch.nih.gov/pdf/>). All these regulations constituted the foundation of similar recommendations in other countries. The first European Union document that specifically relates to protection of medical data is *Recommendation No. R (97)5 of the Committee of Ministers of Member States on the protection of medical data*, https://www.apda.ad/system/files/medical_data_en.pdf.

The HIPAA Privacy Rule provides the following standard for the de-identification of protected health information:

“health information is not individually identifiable if it does not identify an individual and if the covered entity has no reasonable basis to believe it can be used to identify an individual”

There are two levels of document protection. The first is called “Expert Determination” and draws attention to the possibility of identifying a person on the basis of very specific information that is considered safe, quoted below.

“A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

- Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information;
- Documents the methods and results of the analysis that justify such determination.”

For example, the combination of a very rare disease, age and sex may indicate a particular person for experts.

The second method, called “Safe Harbor”, indicates information or part of it that directly identifies a person, and therefore, has to be removed from the data to make it de-identified. 18 identifiers, which apply for “information of the individual or of relatives, employers, or household members of the individual that should be removed from medical documentation”, formulated under the HIPAA Privacy Rule are enumerated below.

1. Names;
2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code, if according to the

- current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
 4. Phone numbers;
 5. Fax numbers;
 6. Electronic mail addresses;
 7. Social Security numbers;
 8. Medical record numbers;
 9. Health plan beneficiary numbers;
 10. Account numbers;
 11. Certificate/license numbers;
 12. Vehicle identifiers and serial numbers, including license plate numbers;
 13. Device identifiers and serial numbers;
 14. Web Universal Resource Locators (URLs);
 15. Internet Protocol (IP) address numbers;
 16. Biometric identifiers, including finger and voice prints;
 17. Full face photographic images and any comparable images;
 18. Any other unique identifying number, characteristic, or code (note this does not mean the unique code assigned by the investigator to code the data).

A short summary of the above regulations and their application to Polish medical data is given by Borucki (2009).

The book *Guide to the De-Identification of Personal Health Information* by Khaled El Emam gives an exhaustive presentation of the problem. The author discusses the risk of disclosing data depending on the way data is made available. He stresses that different approaches are necessary if the data is available on the Internet without any restrictions and if the data is available solely for audited users who can only perform certain operations on the data. The author describes cases when disclosure of personal data is possible by linking data with other available datasets and discusses the method of measuring the risk of disclosure. An estimation of a disclosure risk is important as the cost of a safe de-identification is high; so it is reasonable to perform a de-identification adequate to the risk. Moreover, it is necessary to take into account that data should enable the drawing of correct inferences after de-identification.

The author discusses different methods of de-identification depending on further application of the data. If there is no need to contact a patient in the future,

e.g. to collect additional information concerning him/her, it is possible to suppress some data by substituting it with NULL or just by erasing it. The data might also be substituted by random data that looks real. The last method makes possible to link the same patient's data from different sources of information, if substitution is performed consistently for all sources of course. If it might be necessary to contact a patient in the future, it is recommended that reversible coding is performed, also called reversible pseudonymisation. This might be done using a single code. In this case, there are two datasets: one contains clinical data with personal information substituted with codes, and the second dataset links codes with identification data. The safer approach uses different codes for the two datasets described above, and both codes are linked by the third linking database.

The guide, summarised in brief above, is enlarged upon in another book written by Khaled El Emam and Luk Arbuckle *Anonymizing Health Data*. This time, the authors describe a practical approach to the problem of anonymisation. They introduce two notions: masking and de-identification. The first one is usually done by introducing pseudonyms or by removing the data and is applied to names and information which directly identifies a person. De-identification concerns socioeconomic information like age, race, profession, etc. that should usually be rather generalised or substituted by transformed data to retain its analytic usefulness, instead of being removed. Drawing conclusions should still be possible from less precise data. The authors present methods for treating different types of identifiers and describe possible attacks on the data and discuss risk assessment.

El Emam and Arbuckle concentrate on medical documents in their book, but the same methods might be applied to other domains like insurance, banking, or customers data. One of the important conclusions from this book is: "We can't guarantee zero risk if we want to share any useful data. The very small risk is the trade-off required to realize the many important benefits of sharing health data." It is always necessary to balance possible gains and losses when selecting a method of de-identification, which is never perfect.

4.2 Methods

Methods of constructing tools for automatic medical data de-identification/anonymisation and the problem of their efficiency and evaluation is widely discussed in the literature. The essence of this task is to recognise sensitive information like: names, address, dates, identity codes, etc. in free (unstructured) text. This problem might be seen as a named entity recognition (NER) task. As in other information extraction problems, there are two main methods to solve a NER task: a rule-based or statistical approach.

The rule based approach is currently rarely applied to solve a general NER task but it is quite often applied to medical data (Friedlin and McDonald (2008), Gupta *et al.* (2004), Neamatullah *et al.* (2008)) as it allows us to tune the rules to particular data and has a high performance, and especially, precision.

Unfortunately, such systems require a lot of manual work to develop them and, later, to adapt them to other domains or texts from different sources.

Let us consider an example of the rule based name recognition procedure described in (Neamatullah *et al.*, 2008). The authors use several dictionaries in their de-identification system. These dictionaries contain:

- names of known patients and clinical staff;
- generic first names and last names, hospital names, geographic names;
- keywords or phrases that precede or follow protected information, such as “Mr.”, “Dr.”, “hospital”, “patient”, “age”;
- lists of common words and UMLS (Unified Medical Language System) terms.

The authors use Perl scripts to match regular expressions containing appropriate data from the dictionaries. For ambiguous terms, they apply simple heuristics using context to qualify if information should be removed. In the first step the authors use the most reliable dictionary that contains known names of patients and medical staff and these names are recommended to be removed from texts. In the next step, the algorithm identifies potential names with the help of the list of generic names. These names might be ambiguous as they may match common words or medical terms collected in a dictionary. If they don't match any word in that dictionary, they are marked as protected information and are qualified for removing together with all their occurrences in texts. Otherwise, specific name patterns are checked, such as:

- <first name> <last name>
- <last name> , <first name>
- <first name> <middle name> <last name>
- <first name> <initial> <last name>

If one of the first or last names is unambiguous (i.e. it is a name), the algorithm qualifies both as protected information. Then, the algorithm checks all words that match a first name or a word denoting a person (e.g: “Ms”, “Mrs”, “wife”, “father”) and qualifies them as a name. The list of potential names for a document is created. All words matching that list together with names recognised in other documents concerning the same patient are removed as protected information.

The second approach is based on supervised machine learning methods, for example on SVM (Hara, 2006) or CRF (Gardner and Xiong, 2009) methods or a decision tree learning algorithm (Szarvas *et al.*, 2007). These methods recognise information of interest on the basis of a set of features. Usually they are simple, e.g word case, punctuation, digits, special characters, but sometimes morphological features are taken into account. Such programs require a training set of documents annotated with information intended for removal. In some approaches the existing NER models trained on newswire are used as a starting point and then extended by additional medically protected information. These methods have a high performance in recognition information occurring in typical patterns but they tend not to recognise protected information that occurs

in rare or domain specific contexts in the data. Ferrández *et al.* (2012) compare the two above approaches in their paper which is concluded with the following statement:

“...although machine learning de-identification methods are typically more generalizable than rule-based methods, it is sometimes difficult to know precisely why the method committed an error and additional training data is often required when these approaches are applied to a new dataset”

For English, open source systems are available for de-identification like the Identity Management System created within the i2b2 project (<https://www.i2b2.org/software/>, Uzuner *et al.* (2007)) as well as commercial ones, see for example De-Id system (<http://www.de-idata.com/>, Gupta *et al.* (2004)). An overview of systems created in the last 20 years is given in (Meystre *et al.*, 2010). The problem of anonymisation or de-identification has also been undertaken for other languages: Swedish (Dalianis and Velupillai, 2010), French (Grouin and Névéol, 2014) or German (Onken *et al.* (2009), Tomanek *et al.* (2012)).

Construction of medical corpora in Polish

In the chapter, we describe construction of medical corpora in Polish on the basis of two sets of data consisting of hospital documents. One contains the discharge documents of diabetic patients gathered in a hospital between the years 2001 and 2006. The corpus consists of 460 hospital discharge reports. We described the construction of this corpus in (Marciniak and Mykowiecka, 2011a), while details of morphological analysis were included in (Marciniak and Mykowiecka, 2011b). Initially, we collected the data in order to perform experiments with developing rule based information extraction systems, see (Marciniak and Mykowiecka, 2007) and (Mykowiecka *et al.*, 2009). On the basis of the results of the rule based IE system, we prepared a semantic annotation of the data which we used in experiments with the machine learning information extraction approach based on Conditional Random Fields, see (Mykowiecka and Marciniak, 2011a). The second corpus, we collected from six wards of a children's hospital. It consists of over 1200 discharge records of patients from 2006 – 2007. We used the data to test terminology extraction methods described in (Marciniak and Mykowiecka, 2014b) and (Marciniak and Mykowiecka, 2015) and also in Chapter 8.

We discuss our approach to constructing medical corpora of hospital documentation in Polish. We start from the data anonymisation problem, then shortly afterwards indicate difficulties in converting data into plain text, and exhaustively discuss problems connected with morphosyntactic analysis of texts. Although we concentrate on medical data processing, the same methods can be applied to other domain data.

5.1 Anonymisation

For Polish, the problem of data anonymisation was raised in the context of police reports (Graliński *et al.*, 2009). The authors developed a rule-based named entity recognition formalism used for machine translation and anonymisation. To recognise named entities in Polish texts, two tools based on CRF are available: NERF (Savary and Waszczuk, 2012) and Liner2 (Marciniak *et al.*, 2013). The reported average F-score of the tools is 77.3 and 79.6, respectively. They may support the development of a system for the anonymisation of medical data but require additional adjustments for this task.

Sensitive data can also be removed manually by a person authorised to have access to patient data. This would, however, require many hours of work. Generally, the effectiveness of manual work is deemed to be very high. But, while

processing hundreds of files, it is possible that some data will be left in its original form and sensitive data remains in the file, or will be made only partially anonymous. Additionally, in the case of reversible anonymisation, manual creation of a file with codes used to link a patient with a clinical record is difficult and prone to mistakes.

Anonymisation of the first data, i.e. diabetic patient documents, was manually done in the hospital and identification information was permanently removed. This data was collected to extract complex information in order to draw conclusions concerning diabetes patients, e.g. a correlation of exam results and long-lasting complications. As anonymisation was done by erasing sensitive information from texts, it was not possible to link discharge records concerning the same patient.

For the second set of documents, we prepared a program for reversible anonymisation of patients. We used a rule based approach to recognise sensitive data in documents. The method was described in (Marciniak *et al.*, 2010). Below, we recapitulate the problems (and solutions) that we came across during the development of the program, especially in the key-code file construction. The program substitutes identification information in documents with a patient's code and creates a file that allows us to link the code with the identification data. It is worth noting that the same file should be used every time when new data is anonymised in order to apply the same code in several documents about the same patient. To recognise personal information we use patterns that indicate introductory phrases in texts. For example, for address recognition we use the following phrases: *Miejsce zamieszkania* 'Place of residence' *Adres zamieszkania* 'Address of residence' or *Zamieszkały* 'Residing' together with a phrase that follows it. Moreover, files that contain discharge documents are labelled with the patient's name, so these filenames also have to be changed. Our program uses codes as file names instead of patient's names.

The key-code file should be constructed in a way that allows us to link all the discharge documents of the same patient. The unambiguous method for identifying a citizen in Poland is to use his/her PESEL number. After the inspection of a sample of discharge documents with fictitious identification data it turned out that not all documents contained a PESEL. The substantial problem was that in hospitals which didn't use any professional program for hospital service, discharge documents were not standardised and contained various patients' identification information edited in a number of different ways. The sample data showed that all documents contained the surname, forename and address of a patient. The PESEL was usually present but sometimes only the date of birth was given. As far as children's hospitals with wards for newborn infants are concerned, the use of PESEL as the patient identification code is irrelevant (they do not have this number assigned yet).

The method of key-code creation must be homogeneous for all documents. As it happened that the same patient obtained a discharge document with a PESEL and without it during two successive hospital visits, we gave up identifying a patient by PESEL. We decided that the method of a patient's identification

by the surname, forename and date of birth was the best we could use, even though, theoretically, it was not perfect. So, our key-code file contained the following information: surname.firstname.birthdate with a unique code starting with M for males and F for females. For example, a code file contained the line:

```
KOWALSKI_JAN_05022001 M080001
```

which means that Jan Kowalski born on 05.02.2001 was represented in documents by the code M080001.

There are several problems with normalisation of surnames and forenames used as a part of the identification code. To allow for unification of strings: Kowalski Jan and KOWALSKI Jan we convert all of them into capital letters. It is also necessary to take into account the order in which surname and forename occur. In most documents a surname precedes a forename but in some reports, they appear in the reverse order, so the program should code: Jan Kowalski and Kowalski Jan as the same string: KOWALSKI_JAN. We apply simple heuristics to establish the order of a surname and a name in a document. First, we use an opening phrase: *Imię i nazwisko* ‘Surname and forename’ or *Nazwisko i imię* ‘Forename and surname’ if it is present. Then, we use a list of forenames and determine which string is on the list. This solution will not work out if the surname of a person can also be interpreted as a forename. If the above methods fail, we use the same order as in the most recently processed document.

The birth date is the crucial information that needs to be normalised by the program to create a patient code. We have to identify that the following strings: ‘5.02.2001’, ‘05.02.01’, ‘05.02.2001’, ‘05-02-2001’, ‘05. 02 .2001’ relate to the same date and are assigned to the same identification string: 05022001. In documents all birth dates are written in the following order: day-month-year, but we also take the reverse order into account.

The program was tested on 300 documents from 6 wards (50 documents from each one). It successfully processed 96% of the documents and all patient identification data was removed. The documents were additionally scanned to verify that the removed names were not present in texts. The program abandoned processing a document if it encountered any problem. So five documents were not processed because we did not predict all data formats, e.g. our program did not accept months written as roman numerals: ‘05.II.2001’. Three documents from one ward were not patient discharge reports, and another two documents did not contain the crucial data necessary for the key-code file. Two documents were not processed because they were duplicates of already processed documents.

The described method of de-identification is not sufficient to make the medical data available as it removes only information that directly identifies a person. Moreover, the collected data might be used only for purposes specified in the contract between the Institute and the medical centres.

Finally, let us add that the collected discharge records came from 2001–2007, when hospital documents were written in text editors as, at that time, programs for conducting electronic medical records were not available in the hospitals

we collaborated with. Currently, most Polish hospitals use electronic systems¹ so problems concerning anonymisation have changed a bit. But still, historical data needs an approach similar to that described in this section.

5.2 Conversion into plain text

Domain data are collected from different sources in different formats like: pdf, HTML, MS Word. These formats have to be converted into plain text. There are several publicly available or commercial tools that can perform this task. Some errors that are introduced by such programs are easy to correct. For example, tools converting pdf files into plain text quite often don't handle properly hyphenated words. In such cases, it is necessary to postprocess the results of conversion in order to join both parts of the words. For medical data, if we want to reflect all nuances of the content, it is more useful to prepare a dedicated tool.

Hospital records have to be converted into plain text with a thorough accuracy. It is required to pay attention to text formatting, for example to represent tables in a way that makes it possible to reproduce its content. It is necessary to represent information in subscripts properly (e.g. H_2O) and superscripts (e.g. 10^2). If we ignore the information represented by the text format we get strings: $H2O$, 102 for the above examples. In the first case, a program using data may recognise common errors in chemical formulae. Unfortunately, the later example shows that if we ignore the superscript format, we obtain plausible numerical information and we have no evidence suggesting that this value might be incorrect.

The collected hospital documents were originally written in different versions of the MS Word editor. Some of them were written in the Open Office Writer editor. To take into account several formats and issues raised above, we prepared our own tool for converting them into plain text.

5.3 Document structure

Hospital discharge summaries usually have similar structure and provide the same type of information. Despite an established scheme of documents, they differ significantly depending on hospitals, requirements of a particular ward, and the writing style of a doctor.

Each document has a heading that contains information concerning a hospital, e.g. its name, address, telephone number, ward, and sometimes logo. It may happen that this information is contained within an image. Just after a header, information identifying a hospital visit is given: the unique number within a year, the date of the document, the place where it was written, and the dates of admission to hospital and leaving it. Each document starts with the same fixed

¹ In accordance with the *Act on the medical information system* in Poland, all hospitals and clinics should run medical documentation in electronic form from 1 August 2017.

phrase: *Karta informacyjna* ‘Information card’ or the longer one: *Karta informacyjna leczenia szpitalnego* ‘Hospital treatment information card’ followed by personal (protected) patient information. After introductory information, there are usually blocks of texts containing the following data:

Diagnosis of a disease, due to which the patient was hospitalised.

Test results are presented in the form of text, e.g. description of ultrasound or X-ray exams; as numbers, e.g. height, weight; or in tables, e.g. blood tests or a lipid profile.

Treatment contains short information on both medications administered during hospitalisation and therapeutic procedures.

Discharge abstract contains an analytical summary of the causes and course of the disease; repeats the most important test results and gives an overview of treatment.

Recommendations describe treatment after leaving hospital, e.g. what medications the patient should take and what kind of diet should be observed, what tests he/she should perform, and where to continue the treatment if necessary.

Each document, from both hospitals, is 1–2.5 pages long after conversion into plain text.

5.4 Linguistic characteristics of data

The vocabulary of clinical documents is very specific and significantly differs from general Polish texts. Therefore many medical words are not present in Polish general dictionaries.

Medical texts contain a lot of medication names, like *Cefepime* or *Furagin*. Some of them are multi-word names, like: *Diaprel MR*, *Mono Mack Depot*, *Mixtard 10*. There are several ways to write the same medication depending on international or Polish spelling rules (e.g. *Amitriptylinum* and its Polish equivalent *Amitriptylina*). Moreover, Polish names can be inflected by cases, e.g. *Amitriptyliny_{gen}* or *Amitriptyliny_{inst}*.

Many diagnoses are written in Latin, e.g. *immobilisatio gypsea* ‘immobilization with gypsum’. Sometimes, a whole phrase is in Latin: *Retinopathia diabetica simplex cum maculopathia oc. sin.* ‘simple diabetic retinopathy with maculopathy of the left eye’, or *Laryngitis chronica. Otitis media purulenta chronica dex.* ‘Chronic laryngitis. Chronic purulent inflammation of the middle right ear’. Foreign expressions can be thrown into Polish sentences: *Ascites duża ilość płynu w jamie brzusznej* ‘Ascites a lot of fluid in abdominal cavity’ — only the first word is not in Polish. Bacterial names are frequently written in Latin: *Pseudomonas aeruginosa*, *Streptococcus agalactiae* or *Staphylococcus aureus*.

Medical documents contain many abbreviations and acronyms, some of them are in common use: *TK* ‘CT’ (Computed Tomography) or *godz – godzina* ‘hour’, but many of them are domain dependent. For example, *por.* in everyday language

means *porównaj* ‘compare’, but in the medical documents it abbreviates: *poradnia* ‘clinic’. Units in results of exams are given in abbreviated forms like: *min* – ‘minute’, *mm* – ‘millimetre’ or *tab* – *tabletki* ‘pillow’. Some abbreviations are created ad hoc, e.g., in the phrase *babka lancetowata* ‘ribwort plantain’ the word *lancetowata* ‘ribwort’ is abbreviated to *lan* or *lanc*. In *j. brzusznej* (‘abdominal cavity’), *j.* is the abbreviation of *jama* ‘cavity’. These abbreviations might only be properly interpreted in context as they are not included in any dictionary of abbreviations.

A lot of information is represented in numerical form, e.g. test results with numerical values, dates (in different formats), time intervals, frequency of events (e.g. how many times per day a drug should be given), and dosages. The description of digital strings is usually omitted in general corpora. For example, in NKJP, all such strings have an *ign* (ignored) description. But, in medical documents, numerical values contain very important information that should be properly interpreted. For example: *RR: 120/70 mm Hg*, where *RR* is the acronym denoting the method of a blood pressure examination, while *mm*, *Hg* denotes units.

As medical records are not meant to be published, they are not very carefully edited. Despite the spelling correction tool being turned on, some errors still occurred. The majority of errors are in words that are not included in a standard editor dictionary, like *elaktroresekcji* instead of *elektroresekcji* ‘electroresection’_{gen} but in common words spelling errors are also quite frequent. A typical error is the lack of Polish diacritics, e.g. *miesiace_{nom}* or *miesiecy_{gen}* instead of *miesiące* or *miesiący* ‘months’, as introducing diacritics requires the use of additional key.

5.5 Tokenisation

The first level of every linguistic analysis of a text is its division into elementary components, i.e. tokens. Tokens are, roughly speaking: words, abbreviations, punctuation, and numbers. Usually, not much attention is paid to tokenisation and texts are just divided at spaces, punctuations, or ends of lines. In many applications, such a solution can be sufficient. Most tokens in novels are: lowercase words, words beginning with a capital letter, and punctuation marks. The most important tokenisation problem is then to decide whether a particular dot ends a sentence or belongs to the preceding abbreviation (or both). For texts containing a lot of information other than words, like numerical data, units or formulas, such a division is highly insufficient. But even for general texts, tokenisation affects the results of text analysis. Dridan and Oepen (2012) compare three tokenisation methods performed on general English texts: Penn Treebank (PTB) (Santorini, 1990), Stanford CoreNLP <http://nlp.stanford.edu/software/corenlp.shtml> and tokenisation done by a parser described in (Charniak and Johnson, 2005). They indicate following sources of differences:

“under-restricted punctuation rule that incorrectly splits on commas within numbers or ampersands within names. Other than that, the prob-

lematic cases are mostly shared across tokenisation methods, and include issues with currencies, Irish names, hyphenization, and quote disambiguation.”

Let us consider Examples 5.1–5.4 containing fragments of discharge records. They illustrate problems related to the tokenisation of medical texts, especially hospital documents.

(5.1) *HbA1C:10.6 % (norma 4,2 -5,7)*

HbA1C:10.6 % (norm 4,2 -5,7)

(5.2) *Badanie ogólne moczu: Zabarwienie żółte, odczyn pH 6.5, ciężar właściwy 1.010, białko nb, cukier nb, aceton nb, urobilinogen 0.2 E.U./dl, leukocyty 3 - 5 w.p.w., erytrocyty 2 - 4 w.p.w.,*

General urine examination: Yellow colour, pH 6.5, specific weight 1.010, protein abs(ent), glucose abs, acetone abs, urobilinogen 0.2 E.U./dl, leucocytes 3 - 5 within eye-shot, erythrocytes 2 - 4 within eye-shot,

(5.3) *Badania biochemiczne:*

Mocznik 59 mg/dl, kreatynina 2,7 2,4 mg/dl, klirens kreatyniny D-54 ml/min., N -47 ml/min, Na+ 136 138 mmol/l, K+ 9,3 4,6 mmol/l, Ca 2,1 mmol/l, ferrytyna 140 ng/ml, AST 22 u/l, ALT 66 u/l, ALP 151 u/l, GGT 39 u/l, troponina 0,06 ng/ml, czas PT 12,9 sek., Wskaźnik protrombiny 82,9%, INR +1,25;.

Biochemical analysis:

Urea 59 mg/dl, creatinine 2,7 2,4 mg/dl, creatinine clearance D-54 ml/min., N -47 ml/min, Na+ 136 138 mmol/l, K+ 9,3 4,6 mmol/l, Ca 2,1 mmol/l, ferritin 140 ng/ml, AST 22 u/l, ALT 66 u/l, ALP 151 u/l, GGT 39 u/l, troponin 0,06 ng/ml, PT 12,9 sek., Prothrombin ratio 82,9%, INR +1,25;.

(5.4) *Dieta cukrzycowa 2100 kcal, 4 posiłki/dobę (3 posiłki główne + II kolacja).*

Insulina:

R 34j. Mixtard 50

P 6j. Actrapid

W 20 j. Mixtard50

Systematyczne przyjmowanie leków:

Glucobay 50+50+0

Xartan 1x1/2 tabl.

Acard 1x 1 tabl.

Digoxin 1x1 tab. przez 5 dni 2 dni przerwy

Diabetic diet 2100 kcal, 4 meals/day

(3 main meals + II supper).

Insulin:

M(orning) 34j. Mixtard 50
M(idday) 6j. Actrapid
E(vning) 20 j. Mixtard50
Systematically taking medication:
Glucobay 50+50+0
Xartan 1x1/2 tabl.
Acard 1x 1 tabl.
Digoxin 1x1 tab. through 5 days 2 days break

As can be seen, a lot of relevant information is given in the numerical form. Apart from time descriptions (dates, hours, periods of time), there are numbers that refer to values of medical tests or medicine doses, amounts and sizes. Moreover, physicians do not pay enough attention to punctuation rules. For example, they use comas and dots interchangeably to write decimal numbers so they mix Polish and English standards. Moreover, there are specific names which contain non-letter characters, like $Na+$ for a sodium cation or $K+$ for a potassium cation.

In the rest of this section, we describe the tokenisation method applied to diabetic corpus construction. As all annotation levels refer to tokens, they have to be granular enough to represent other annotations, i.e. the morphological annotation and the semantic annotation. The first annotation is done with the help of the TaKIPI tagger (Piasecki, 2007) while the second one is prepared on the basis of the SProUT grammar, see Chapter 6. So we have to take into account both tokenisations applied in TaKIPI and SProUT.

The general assumption adopted in the SProUT tokeniser is “not to interpret too much”, which means that tokens are relatively simple and do not rely on any semantic interpretation. Their self-explanatory names, together with token examples are listed in Table 5.1.

The TaKIPI tagger provides two tokenisation methods. The simple one divides all character sequences into words and non-words. A non-word is any sequence of characters containing non-letters, but not containing: spaces, line breaks, or more than one of any punctuation character except a hyphen. Non-words have the *ign* label assigned. The second TaKIPI tokenisation method creates tokens too complex to be taken into account in our task.²

We decided to use the token classes identified by the SProUT tokeniser and to align its results with the results of the simple TaKIPI tokeniser. SProUT tokens which were longer than TaKIPI tokens, e.g. ‘*1x2mg*’, ‘*100mg*’, ‘*50x16x18*’, were divided into smaller ones.

The changes introduced to token limits concerned those SProUT tokens of the *other_symbol* type which contained punctuation marks. The *other_symbol* class comprised sequences which did not fit into any other class, i.e. symbols for which separate classes were not defined (e.g. ‘=’) and mixed sequences of letters and digits. In this latter case a token ended only when a space or a line break

² For all non-word tokens, only a space or a line break ends a token so, for example, sequences ‘*200/min.*’, ‘*24,9%,bilirubina*’, ‘*25.3-61,5%*’, ‘*dnia13/14.07.04*’, ‘*iVS-1,5*’, ‘*ml/h-3*’ are singular tokens.

Table 5.1: Token types and number of occurrences

Token class name & Examples	Numbers	
	initial	final
<i>all_capital_word</i> : ALT, B, HDL, HM	18,369	18,416
<i>any_natural_number</i>	85,766	87,246
<i>apostrophe</i>	14	14
<i>back_slash</i>	7	7
<i>closing_bracket</i>	2,661	2,663
<i>colon</i>	12,426	12,427
<i>comma</i>	28,799	28,831
<i>dot</i>	47,261	47,269
<i>exclamation_sign</i>	49	49
<i>first_capital_word</i> : Al, Amikacin, Wysokie	43,136	43,269
<i>hyphen</i>	4,720	4,725
<i>lowercase_word</i> : antygen, aorta	192,305	193,368
<i>mixed_word_first_capital</i> : AgHBs, Ilo, NovoRapid	513	514
<i>mixed_word_first_lower</i> : antyHBS, dIAST	989	1,003
<i>number_word_first_capital</i> : 200Hz, 14HN	48	0
<i>number_word_first_lower</i> : 100ml, 200r 1kaps	650	0
<i>opening_bracket</i>	3,344	3,355
<i>other_symbol</i> : (132x60mm), 1,34x3,25, HbA1c=10,3, ml/min.	3,161	2,868
<i>percentage_tok</i>	4,461	4,478
<i>question_mark</i>	207	209
<i>quotation</i>	1	1
<i>semicolon</i>	455	455
<i>slash</i>	10,340	10,353
<i>word_number_first_capital</i> : AST34, B6	1,195	1,195
<i>word_number_first_lower</i> : mm3, pH6	1,865	1,854
<i>word_with_hyphen_first_capital</i> : B-hCG, Anty-HBs	163	163
<i>word_with_hyphen_first_lower</i> : m-ce, p-cial	402	402
All tokens	463,307	465,004

was encountered. The common case when this strategy failed in our data was the sequence ‘*HbA1c:*’ as the name of the test HbA1c and the following colon was classified as an *other_symbol*.³ To make the results more uniform we divided these tokens at the point where punctuation characters occur. Among these newly created tokens the most numerous class was *lowercase_word* and numbers which were formed after separating numbers and unit names, e.g. *10g*, *100cm* and sequences describing repetitions or sizes, like *2x3*, *2mmx5mm*. Sometimes we divided text into smaller fragments than TaKIPI. This happened mostly when

³ Examples of the other similar sequences: ‘*HbA1c=9,1%:*’ or ‘*(HbA1C)*’.

the tagger did not behave uniformly – it usually divided the number and ‘ml’ string into two tokens, but not always.

Table 5.1 gives the comparison of occurrences of token types in the initial and final approach. Finally, in the entire diabetic data corpus 465,004 tokens are identified, numbers represented 18.8% (9% of characters), and all punctuation characters constituted 25% of the total number of tokens (6.5% characters).

5.6 Morphological analyses

Morphological analysis is an important stage in any language processing task. For each word form it gives information about its lemma, part of speech and its grammatical features. It is especially important for highly inflectional languages, like Polish, for which the grammatical description of words is much more complex than for English.

In this chapter we describe the morphological analysis of discharge documents of diabetic patients, previously reported in (Marciniak and Mykowiecka, 2011b). Morphological annotation of these texts is based on the results obtained by the publicly available Polish POS tagger TaKIPI that cooperates with *Morfeusz* SIAT Woliński (2006)⁴ — a general-purpose morphological analyser of Polish. For each word form, *Morfeusz* assigns all possible interpretations containing: its base form, part of speech, and complete morphological characterisation (e.g. case, gender, number, aspect if relevant). The description is exhaustive and aimed at further syntactic analyses of texts.

The documents were analysed and disambiguated by TaKIPI. TaKIPI can be combined with the *Guesser* module (Piasecki and Radziszewski, 2007) which suggests tags for words not included in the dictionary. We decided to use this module because, otherwise, 70,600 tokens representing words and acronyms that occurred in the documents would be assigned an unknown description. The gain from its usage was, however, not so evident, as tags and base forms suggested by *Guesser* were quite often incorrect – in one test set, only 272 forms out of 1,345 were analysed correctly.

The analysis of TaKIPI results shows that there are many systematic errors. An example of such an error is the description of medication names produced by *Guesser*. Their morphological tags are often correct, but the problem is with gender assignment in the case of masculine forms. In Polish there are three subtypes of masculine gender: personal, animate and inanimate, and *Guesser* quite often uses the personal masculine gender instead of the inanimate one while analysing medication names. The second most common problem concerns base forms, because all base forms created by the module are written with a small letter. So for proper names, all base forms have to be corrected. TaKIPI

⁴ The new version of *Morfeusz* was released in 2014, see Woliński (2014). New functionalities enable a domain vocabulary to be added to the main dictionary, to analyse compound words and to disallow some interpretations, e.g. archaisms.

does not disambiguate all tags – certain forms still have more than one possible description left which should be disambiguated.

The results of TaKIPI were postprocessed with a set of global correction rules (see Section 5.6.2). The rules mainly corrected the description of:

- acronyms and units: *BMI*, *HbA1c*, *RR*, *USG*, *Hz* or *kcal*;
- medication names, for *Diaprel*, the *diaprel* base form should be changed into *Diaprel*;
- domain terms like *dekarboksylazie* (‘decarboxylase’_{loc}) for which the masculine base form *dekarboksylaz* was suggested instead of the feminine *dekarboksylaza*;
- misspelled tokens;
- foreign words;
- if there was more than one description attached to a word form, then the more probable one in the domain was chosen.

After applying the global correction rules a fragment of data was manually corrected to evaluate the effectiveness of the method.

5.6.1 Tags

Polish is a language with rich inflection (7 cases, 2 persons, 5 genders). The Polish tagset (see Appendix A) is quite detailed; the set of potential morphological tags consists of more than 4,000 elements. In the general corpus of Polish over 1,000 tags are used for around 30 word classes (reported by Przepiórkowski (2005) for the IPIPAN corpus), while in the diabetic corpus only 450 different tags are represented. For example, there are no verbs in the first and second persons, or in a conditional mood or any words in a vocative case.

Despite the very detailed tagset we extend it by several new categories that allows us to describe errors, abbreviations and foreign words more precisely. If no tag suits a token still, the tag **tsym** is assigned to it. In particular, all patient codes (like *d2005_006*) have the **tsym** tag.

Errors

Spelling errors in the corpus are left as they are, so in the final data they are not directly corrected. There are no changes of tokenisation introduced to the data in the case of lack of space or additional space. Misspelled tokens are assigned the base form equal to the token, and one of the following tags, depending on the type of error:

- **err_spell** describes misspelled tokens like *bia3ko* instead of *białko* (‘protein’). In the corpus, we allow for introducing additional information with the corrected input token, its base form and morphological tag. So, it is possible to reproduce the corrected text of document.

- **err_conj** describes concatenations like *cukrzycowej2000* ('diabetic2000'). In this case we add the correct form *cukrzycowej 2000* to the corpus but we don't introduce any changes in tokenisation and don't add any morphological interpretation of the corrected forms.
- **err_disj_f** describes the first part of an incorrectly disjointed word. For example, if the word *ciśnienie* 'pressure' is divided into two parts *ci* and *śnienie*, (by chance, both are valid Polish words), the first part *ci* is labelled by the tag.
- **err_disj_r** describes the second (or subsequent) part of the incorrectly disjointed word.

The last three categories can be supplemented with **spell** description, if necessary. For example, the token *Bylaw* is a concatenation of the misspelled word *Była* ('was') with the preposition *w* ('in'). This token has the **err_conj_spell** tag, and the *Była w* correction is added.

Abbreviations

There are many abbreviations in the documents. Some of them are used in general Polish like *prof.* ('professor') or *dr* ('doctor'), but there are many abbreviations that are specific to the medical domain. For example, in the descriptions of USG examinations the letter *t* denotes *tętnica* ('artery'), while *tt* refers to the same word in plural. Usually the same abbreviation can be used in plural and singular contexts, e.g. *wit* ('vitamin'). Sometimes it is not a single word but the whole phrase which is abbreviated, e.g. *NLPZ* is the acronym of the noun phrase *Niesterydowe Leki PrzeciwZapalne* 'Non-Steroidal Anti-Inflammatory Drugs', and *wpw* is the abbreviation of the prepositional phrase *w polu widzenia* 'within eye-shot'. Abbreviations and acronyms obtain the **acron** tag. Moreover, it is possible to insert the full form corresponding to them.

Acronyms denoting units obtain the **unit** tag. Units in common usage are not explained: *mm*, *kg*, *h*, but if a unit is typical to the medical domain, its full form can be added (e.g. *HBD* means *tydzień ciąży* 'week of pregnancy').

We also distinguish two tags describing prefixes and suffixes. The token *makro* ('macro') in the phrase *makro i mikroangiopatia* ('macro and microangiopathy') has the **prefix** tag, while the **suffix** tag describes, for example, the part *ma* of the string *10-ma* which indicates the instrumental case of the number 10, like in: *cukrzyca rozpoznana przed 10-ma laty* ('diabetes diagnosed 10 years ago').

Foreign Words

Foreign words receive the **foreign** tag. This tag can have additional information on the part of speech. For example, *Acne* has the **foreign_subst** tag description while *minoris* has the **foreign_adv** tag.

5.6.2 Global correction rules

As the tools we used for morphological analysis (the dictionary and tagger) were not specifically tailored for processing medical texts, the results we obtained were far from our expectations. So we decided to correct systematic tagging errors and Guesser's suggestions. We manually prepared a set of global correction rules that worked without context. They only eliminated evident errors. Correction rules have been created on the basis of the frequency list of all different token descriptions. Each rule was applied to all matching token descriptions in the already tagged data.

The method of global changes allows us to reduce the number of manual corrections in the corpus at the final, manual stage. But, it should be noted that rules operating without context are not able to properly correct all mistakes. We can only eliminate evident errors but we cannot decide, for example, if the description of the adjective *ciężkim* 'heavy' is correct or not in a particular usage. In the frequency list of all different token descriptions there are the following occurrences:

- 1# *ciężkim*#*ciężki*#*adj:sg:inst:m3:pos*#
- 8# *ciężkim*#*ciężki*#*adj:sg:loc:m3:pos*#
- 2# *ciężkim*#*ciężki*#*adj:sg:loc:n:pos*#

The initial number informs us about the frequency of the interpretation of *ciężkim* 'heavy' as a form of the lemma *ciężki* with the tag describing a singular (*sg*) adjective (*adj*) of impersonal masculine gender (*m3*) or neutral gender (*n*) in *inst* – instrumental or in *loc* – locative case and *pos* – positive degree. The form *ciężkim* may also have several other interpretations, e.g. as an adjective in plural. All these tags may only be verified if we know the noun which is modified by the adjective, i.e. the context in which the adjective occurred.

However, quite a lot of corrections may be done in any context, e.g. changes of gender of a medication name (*Lorinden_f* into *Lorinden_{m3}*), or in the prevailing number of cases, e.g. assigning to *zwolnienie* the *gerund* tag 'slowing' (11 occurrences) instead of the less frequently occurring *noun* 'sick leave' (only one occurrence; TaKIPI leaves both descriptions).

There are two types of correction rules, of which syntax is given in 5.5–5.6 where: '#' is a separator; the character '>' indicates the new token description that is applied to the corpus; after || additional information can be noted. In the case of Rule 5.5 it could be a text fragment that explains the meaning of an acronym, abbreviation or a foreign word, while for Rule 5.6, a corrected token, its base form and tag can be given. This additional information might be used for creating a corpus without spelling errors, or for creating dictionaries of abbreviations or foreign words used in the medical domain.

- (5.5) token#base form#tag#>
 token#new base form#new tag# || 'string' (optionally)
- (5.6) token#base form#tag#>
 token#token#error_spell# || corr. token#corr. base form#new tag#

Rule 5.5 is useful for changing the base form or the tag of a token. See Rule 5.7 where the first letter of the base form is capitalised, the POS of the token is changed from the adjective into the substantive. The description only contains grammatical categories appropriate to nouns, so they have no degree. The personal masculine gender *m1* is changed into the inanimate masculine gender *m3*.

(5.7) Gopten#gopten#adj:sg:nom:m1:pos#>
Gopten#Gopten#subst:sg:nom:m3#

Rule 5.6 is applied to a token *graniach* ‘ridges’ (in a mountain) that represents the existing but unreliable word in the medical domain. For all of its occurrences in our data (3 cases) it should be substituted by *granicach* ‘limits’. It might be done by the following correction rule:

(5.8) graniach#grań#subst:pl:loc:f#>
granicach#granicach#err_spell# ||
granicach#granica#subst:pl:loc:f#

If there is more than one interpretation left by TaKIPI, all are mentioned before the character ‘>’. See Rule 5.9 where two different base forms are possible for the token *barku* and both have the same tag assigned. The second base form *bark* (‘shoulder’) is definitely more probable in the medical domain than the first one *barek* (‘small bar’ or ‘cocktail cabinet’), so the rule chooses the first description.

(5.9) barku#barek#subst:sg:gen:m3##bark#subst:sg:gen:m3#>
barku#bark#subst:sg:gen:m3#

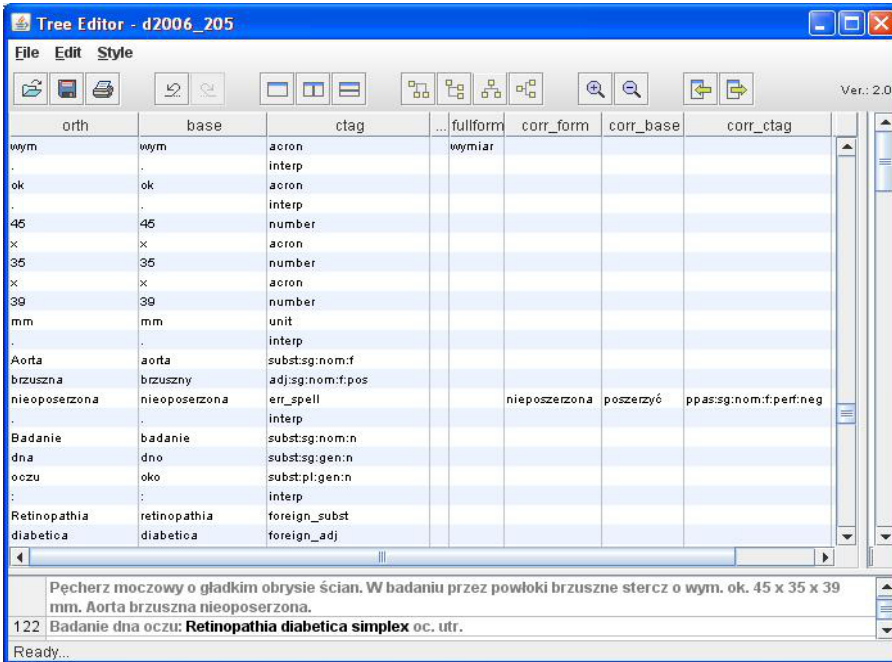
Rule 5.10 illustrates a correction of concatenated token *BMI28* that ought to be written with space as *BMI 28*.

(5.10) BMI28#BMI28#acron#>
BMI28#BMI28#err_conj# || ‘BMI 28’

Figure 5.1 shows a fragment of morphologically annotated text in which new tags are applied. For example ‘mm’ has assigned the *unit* tag and *Retinopathia diabetica* — *foreign_subst* and *foreign_adj* respectively. We can also see an example of the *err_spell* tag usage together with the corrected forms and annotation.

Table 5.2 shows the change in the number of occurrences of the top level morphological classes. It shows their frequency: directly after running the tagger, after changing the token limits and after applying automatic changes. The number of different forms in every POS class in the corpus is given in the last column.

Most POS names are self-explanatory, the full list and description of all morphological tags can be found in (Przeziórkowski, 2004), the newly introduced tags are marked with ‘*’. Of all words (all tags apart from *interp* (interpunction), *number* and *tsym*) the most numerous groups are nouns (*subst*) – 54% and adjectives (*adj*) – 15% of word form occurrences.



orth	base	ctag	...fullform	corr_form	corr_base	corr_ctag
wym	wym	acron	wymiar			
.	.	interp				
ok	ok	acron				
.	.	interp				
45	45	number				
x	x	acron				
35	35	number				
x	x	acron				
39	39	number				
mm	mm	unit				
.	.	interp				
Aorta	aorta	subst:sg:nom:f				
brzuszna	brzuszny	adj:sg:nom:f:pos				
nieoposzerzona	nieoposzerzona	err_spell		nieposzerzona	poszerzyć	ppas:sg:nom:f:perf:neg
.	.	interp				
Badanie	badanie	subst:sg:nom:n				
dna	dno	subst:sg:gen:n				
oczu	oko	subst:pl:gen:n				
.	.	interp				
Retinopathia	retinopathia	foreign_subst				
diabetica	diabetica	foreign_adj				

Pęcherz moczowy o gładkim obrysie ścian. W badaniu przez powłoki brzuszne sterzc o wym. ok. 45 x 35 x 39 mm. Aorta brzuszna nieoposzerzona.

122 Badanie dna oczu: **Retinopathia diabetica simplex** oc. utr.

Ready...

Fig. 5.1: Morphological annotation

If we don't take into account *number*, *tsym* and the punctuation tokens, we obtain 348,461 tokens (TW), out of which 78,854 (29.81%) are changed. The most frequent changes concern introducing domain related *unit* and *acronym* classes (nearly 72% of changes). Quite a number of changes result from the capitalisation of the proper name lemma. In Table 5.3 the numbers of some other types of changes are given.

5.6.3 Evaluation

In this section we present an evaluation of the morphological annotation done on the basis of 8 documents corrected by two annotators. In the case of inconsistent corrections, the opinion of a third annotator was taken into account.

The verified dataset consisted of 8,919 tokens — 4,972 were words, acronyms and units for which verification was essential. The remaining 3,947 tokens represented numbers, punctuation and **tsym** tokens, and they were not subject to verification. The correction rules changed the descriptions of 1,717 (34%) tokens in this dataset. For 87 cases, mainly proper names of medications, the changes were limited to the case of letters, i.e. a lowercase letter was substituted by a capital letter in the base form. Manual verification left 4,497 token descriptions unchanged, while 10.6% of descriptions were modified. To compare

Table 5.2: Morpheme types and numbers of occurrences

POS tag	Number of tag occurrences			Different forms in final corpus
	tagger results	after tok. change	final corpus	
adj	35,305	35,041	36,848	3,576
adv	2,323	2,323	2,437	245
conj	5,852	5,852	5,680	36
prep	29,400	29,400	26,120	71
pron	302	302	142	21
subst	82,215	82,215	105,311	5,093
verb forms:	24,743	24,741	19,912	2,001
fin	2,173	2,173	1,900	190
ger	9,778	9,778	4,677	423
ppas	5,593	5,593	6,170	551
other	7,199	7,197	7,165	837
qub	4,244	4,242	2,452	67
num	703	703	703	34
ign	160,951	163,629	0	0
acron*	0	0	30,003	678
unit*	0	0	28,290	82
prefix*	0	0	13	5
suffix*	0	0	36	6
tsym*	0	0	534	462
interp	115,323	116,556	116,556	21
number*	0	0	87,898	1,386
err_disj *	0	0	179	129
err_spell*	0	0	560	440
foreign*	0	0	1,330	184
Total	461,361	465,004	465,004	14,537

it with results of general text tagging with the TaKIPI tool, Karwańska and Przepiórkowski (2009) report 91.3% accuracy of this task. The Kappa coefficient was equal to 0.983 for part of speech and 0.982 for case assignment (when it is applicable).

The results of manual verification are given in Table 5.4. The ‘Basic tags’ column gives the number of changes of the base form and tag, while the ‘All tags’ column takes into account all changes, including descriptions of the correct word form for spelling errors, explanations of acronyms and units.

More detailed analysis of annotation inconsistencies shows two main sources of errors:

- lack of precision in guidelines resulted in choosing different base forms in the case of spelling errors and different labelling of cases with a lack of diacritics which resulted in correct but not desired forms;

Table 5.3: Morphological tag changes

Type of change	Number	% of changes	% of TW
Base form			
capitalisation only	6,164	13.8	4.12
other	25,503	32.34	9.64
POS			
to acron & unit	56,697	71.90	21.43
to other	10,547	13.37	3.99
Grammatical features (without acron and unit)			
only case	109	0.13	0.04
only gender	1,663	2.11	0.62
other	13,215	16.75	4.99

Table 5.4: Manual correction

	Basic tags	All tags
All tokens	8,919	8,919
Without numbers and interp	4,972	4,972
Unchanged	4,497	4,451
Changed	475	521
same changes accepted	226	228
same changes not accepted	1	1
different changes none accepted	4	5
different changes. accepted 1	3	4
different changes. accepted 2	40	42
only 1st annot. changes - accepted	15	48
only 2nd annot. changes - accepted	128	124
only 1nd annot. changes - not accepted	47	47
only 2nd annot. changes - not accepted	0	0

- some errors were unnoticed by one of the annotators (just a cost of manual work), e.g. in the data there are many strings ‘W’ and ‘w’ which may be either acronyms or prepositions.

There are only a few cases that represent real morphological difficulties, e.g. differentiating adjectives and participles (5 cases among the annotators). Some examples of different case and gender assignments were also observed. They are mostly errors consisting in correcting only one feature instead of two, or a wrong choice of a case for elements of long phrases.

5.7 Extended tokenisation

In the previous section the tokenisation that divides text into simple unstructured fragments is described. This solution makes it easy to analyse any fragment of text, but postpones the interpretation of all complex strings to the next levels of analysis (compare the idea of extended tokenisation by Hassler and Fliedl (2006)). To correctly interpret data in a medical corpus it would be useful to group simple tokens into more complex structures and to assign more precise descriptions to ambiguous punctuation marks.

5.7.1 Complex tokens

Numerical values

We prefer to interpret the whole decimal number as one token, i.e., to assign one description to strings like ‘27,67’. As we noted in Section 5.5, Polish and English standards of coding decimal numbers differ. In Polish we should use commas, while in English dots should be used to separate the fractional part of a decimal number. In informal texts, these rules are not obeyed, so we suggest accepting both styles. Sometimes we observe the inconsistent style of coding decimal numbers even in the same line, see Example 5.1 where the results of the HbA1c test are given with a comma, while normal values are given with dots.

Negative numbers, like ‘-4,5’, are rather rare in medical data, but, if they occur, should be recognised as one complex token. In our data the ‘-’ character seldom denotes a negative number, even if it occurs just before a number. It is often used as a dash, which is sometimes typed without a space between a dash and a positive number (e.g., *waga -67 kg* ‘weigh -67 kg’ or *W badaniach laboratoryjnych hiperglikemia 350 mg%, HbA1c -10,4%* ‘In laboratory tests hyperglycemia 350 mg%, HbA1c -10,4%’). Unfortunately, to interpret these strings correctly, pragmatic knowledge concerning the possible test values is necessary. As the tests mentioned above take almost only positive values, the following strings: ‘-67’ and ‘-10,4’ should be interpreted as two tokens: the dash and the number. Disambiguation of these interpretations requires inspecting available data to check if a negative value is accepted in the context. In our data only one test (concerning arterial blood gas) can have a negative value.

Long numbers are the next problem as they may include spaces which separate three-digit blocs. Typically, in English style, commas are used as three-digit bloc separators. In Polish style, dots are used for this purpose. Unfortunately, in both English and Polish corpora (not only biomedical) spaces are often used instead of commas or dots (e.g., *The initial loan of \$1 500 000 was later increased to \$3 300 000*⁵; *Cztery tygodnie = 40 000 \$*⁶ ‘Four weeks = 40 000 \$’; in our data — *Płytki krwi: 206 000 w mm3* ‘Blood platelets: 206 000 in mm³’.)

⁵ An example from British National Corpus: <http://www.natcorp.ox.ac.uk/>.

⁶ An example from National Corpus of Polish: <http://nkjp.pl/>

Punctuation marks

Punctuation marks are symbols that indicate the structure of information. One punctuation mark can have several functions. For example, a dash can indicate the range of values, relationship, a nested clause or phrase, or itemisation marks. Theoretically, dashes of different length are connected with different functions, but in informal texts they are often mistakenly used, or only one type of dash is used. In the case of our data, a dash can be used in the context of numbers as a minus sign, the value indicator (e.g., *HbA1c - 10,4%*, interchangeably with a colon *HbA1c: 10,4%*), and sometimes as an indicator of a range (e.g., *4-5 posiłków dziennie* ‘4-5 meals per day’), or a proportion (e.g. *Hodowla z rozcieńczenia 10-2* ‘Dilution culture 10-2’). The following fragment *BE 2.0 - -2.7 mmol/l* contains two dashes, the first indicates the range, while the second the negative value.

In many texts, the letter ‘x’ performs the function of a punctuation mark. It has two functions. The first is a size mark, e.g. *torbiele o wym. 12x8 mm* ‘a cyst size 12x8 mm’, where both numbers represent size in millimetres. The second has the ‘how many times’ meaning, e.g. *Glukobay 3x100mg*, where the milligram unit refers to the second number (the amount of medication) and not to the number 3 that indicates how frequently the medication have to be applied.

The next important punctuation mark is ‘/’. One of its functions is to represent a fraction, e.g. *1/2 tabl.* ‘1/2 pill’, where the whole string ‘1/2’ should be recognised as one number. It is also used as a separator of values, e.g. for representation of blood pressure results — *RR: 130/80 mmHg* and for ranges in the case of dates — *02/03.04.2004*.

Dates/Hours

Hours written with dots ‘8.30’ can be mixed up with decimal numbers, while those with a colon can be mixed up with proportions. Disambiguation of these sequences requires analysis of the contexts in which they appear, e.g. *o godz 8.30* ‘at 8.30 hour’ — the word ‘hour’ clearly indicates interpretation.

Units

If we want to interpret any number value, we have to know the unit in which it is given. If we speak about somebody’s age, an implicit unit is the year, but a new born child’s age is measured in months or even days. In our data, a patients’ height is usually given in centimetres but sometimes in metres. To compare information in the following expressions: *160 cm* and *1.6 m* it is required to know the units corresponding to the numbers. The following strings consisting of a sequence of units should be considered as complex units ‘mm/h’, ‘mmol/l’, ‘mg/dl’, ‘mg/l’.

5.7.2 Method

To create complex tokens and interpret ambiguous punctuation we apply a rule based approach. We use the Spejdl tool (see Section 2.1.1) as it is intended

to create complex structures like nominal phrases or complex tokens. Such an approach allows us to take into account the contexts of interpreted strings.

Let us consider several examples of grammar rules when carrying out the task. The rule in Figure 5.2 recognises complex units and creates a complex token from elements described in the `Match` part of the rule. It consists of three to five elements. Some of them have to appear like the first and the last element of the match. They have to have the *unit* type; the middle element is a slash character. Elements might be separated by a space — it is expressed by `ns?`. If we want to refer to the value of any element in the rule result (the `Eval` part), we count all elements described by the rule in `Left`, `Match` and `Right` parts. As a result, an element of the *Compx-unit* type is created. The value of the created element is a string consisting of the orthographic form of the first unit, a slash and the orthographic form of the last unit.

```
Rule "complex-unit-rule"
Match: [pos~"unit"] ns? [orth~"/"] ns? [pos~"unit"];
Eval: word(Compx-unit, 1.orth "/" 5.orth);
```

Fig. 5.2: Rule identifying a complex unit

The rule in Figure 5.3 identifies a unit used to express blood pressure that consists of two consecutive strings ‘mm’ and ‘hg’ or ‘Hg’. These two strings are separated by a space. As a result, a `word` element of the *Compx-unit* type is created with the value `mmHg`.

```
Rule "mmHg-rule"
Match: [orth~"mm"] [orth~"hg|Hg"];
Eval: word(Compx-unit, "mmHg");
```

Fig. 5.3: Rule identifying a unit of a blood pressure result

The rule in Figure 5.4 interprets a slash as a punctuation mark that refers to a blood pressure result — *RR-slash* type of word created in the `Eval` part. The slash is described in the `Match` part with its orthographic form and information that it still has `interp` description. This later information is necessary as the earlier rule might change the slash description into some other interpretation. The `Right` part of the rule describes a context that indicates that the slash is connected with the result of a pressure test. The last element in this context refers to the complex token recognised by the `mmHg-rule` (Figure 5.3). So, it is very important that this complex token is recognised before the `slash-rr-rule` is applied.

The whole grammar consists of 80 rules that recognise complex tokens given in Table 5.5. Many of the rules are domain dependent. Verification of the method was carried out on 10 documents, not used in the process of grammar construction. It gives the following results: the F-score is equal to 98.8 for extended tokens

```

Rule "slash-rr-rule"
Match: [orth~"/" && pos~"interp"];
Right: ns? [pos~"number"] ns? -? ns? [pos~"number"]? ns? [base~"mmHg"];
Eval: word(RR-slash, "/");

```

Fig. 5.4: Rule interpreting a slash in a blood pressure result

Table 5.5: New token types

Token	Examples	Context	Nb of occ.
<i>Date</i>	12.10.2005		2,528
<i>Date-d-m</i>	31.01	31.01/1.03.04, in the day 31.01	12
<i>Date-d</i>	12	12/13.10.2005	394
<i>Date-y</i>	2004	w 2004 r. 'in the year 2004 '	289
<i>Date-m-y</i>	II 2004		8
<i>Hour</i>	12.15, 8:00	o godz 12.15 'at hour 12.15'	7,915
<i>Compx-unit</i>	mmol/l	unit/unit	5,752
<i>Dec-number</i>	23,6 7.25	strings unrecognised as dates or hours	13,939
<i>Numb-space</i>	12 500	values in tables, in text	255
<i>Neg-number</i>	-2.5	BE: -2.5	31
<i>Fraction</i>	1/2	1/2 tab. '1/2 pill'	90
<i>Ion</i>	Na+, K+		666
<i>Proportion</i>	:	blok 4:1 'block 4:1'	10
<i>Range</i>	-, /	norm 2,5 - 4, 12/13.04.2001	2,536
<i>RR-slash</i>	/	RR: 140-155/90-95 mmHg	532
<i>Separator</i>	/	id 0875/10452	4
<i>Schema</i>	-, +	Glucobay 1-0-1, Siofor 850+850+500	58
<i>Times</i>	x	Glucobay 3 x 1 caps.	1,862
<i>Size</i>	x	2x3x5 mm	302

recognition (precision= 99.7%, recall=97,9%). For 13 types of extended token, the F-score is equal to 100. For *Proportion* we did not recognise any case in the data which contained two cases not foreseen by the grammar, so the F-score has an undefined value. The lower F-score (94) of the *Compx-unit* annotation results from errors at the simple tokens annotation level, as the letter *u* was recognised as a preposition instead of a unit. Some problems were also observed in the case of *Range* (95) and *Date* (99).

5.8 Semantic Annotation

As Chapter 6 is devoted to an information extraction task and possible applications of its results, we describe there a method of automatic annotation of corpora with extracted information. Here, we only summarise what kind of information is annotated in the diabetic corpus. We annotate 50 simple attributes,

11 complex structures and 3 lists of structures. They represent the information described below.

- Identification of a patient's visit to hospital: a visit identification number and information if it is a main document or a continuation; the date of the document; dates when the hospitalisation took place.
- Patient information: a patient's identifier, sex, age, height, weight (in numbers or words) and BMI.
- Data about diabetes; if the illness is balanced; when diabetes was first diagnosed (expressed as an absolute or relative date); reasons for hospitalisation (as a list of attributes); and results of basic tests: HbA1c, acetone, LDL, levels of microalbuminuria and creatinine.
- Complications, other illnesses including autoimmunology and accompanying illnesses, which may be correlated with diabetes.
- Diabetes treatment described by an insulin type and its doses; description of continuous insulin infusion therapy; description of oral medications; information that insulin therapy was started.
- Diet description that contains information on type of diet, and structures describing how many calories are recommended and a similar structure representing numbers of meals.
- Information on therapy given in text form, e.g. if a patient obtained important information on the diabetes treatment (education) or if a patient observes the diet and self-monitors the blood glucose, if the therapy is modified.

5.9 Corpus structure

Texts represented in corpora are usually provided with additional information which can be of different types. Texts might be annotated with various levels of information like: morphologic, syntactic or semantic. A good practice in corpora creation is to use established standards of storing and annotating data. In this section we give a short sketch of the medical corpus consisting of diabetic records.

The corpus format is developed on the basis of the format accepted for the NKJP corpus (Bański and Przepiórkowski, 2009) that follows the TEI P5 guidelines. These guidelines are advised for annotation of biomedical corpora too, see (Erjavec *et al.*, 2003). According to this scheme, various annotations are encoded in separate annotation layers. So, each document is represented by a directory containing a set of files encoding particular information. The diabetic corpus contains information appropriate for the following five files, see (Marciniak and Mykowiecka, 2011a) for a detailed description of a file structure:

- *xxx.txt* – plain text of the original anonymised document;
- *xxx.xml* – text of the document divided into numbered sections which are in turn divided into paragraphs;
- *xxx_segm.xml* – token limits and types;

- *xxx_morph.xml* – morphological information (lemmas and morphological feature values);
- *xxx_sem.xml* – semantic labels and limits.

The file *xxx.xml* includes text divided into parts and paragraphs (every line break begins a new paragraph). As the general structure of discharge records is regular, a document is automatically divided into six parts on the basis of introductory phrases, i.e. *Introduction*, *Diagnosis*, *Examinations results*, *Treatment*, *Discharge record* and *Sign*. For this purpose, we use regular expressions. They recognise appropriate introductory phrases and allow for some diversities within them. For example, *wypisany do domu z zaleceniami* ‘discharged home with recommendations’ is a typical phrase that starts the *Recommendation* part of a document. It may contain additional information, e.g. about the condition of the discharged patient: *wypisany w stanie dobrym do domu z zaleceniami* ‘discharged home in good condition with recommendations’. If we limit our recognition of the beginning of the *Recommendation* part to the word *zalecenia* ‘recommendations’, we obtain many false beginnings as this word is used in other parts of the documents for various purposes. Thus, we recognise the coexistence of a word indicating a patient’s discharging like *wypisać* ‘discharge’ with *zalecenie* ‘recommendation’ in a distance of a few words. For 460 documents, this simple algorithm only failed in several cases where the beginning of the recommendation section was marked by a slightly different phrase at the end of a sentence.

Some parts might be omitted in some documents. For example, the following sentence *Zalecenia co do dalszego postępowania pacjent otrzymał od konsultującego lekarza.* ‘The patient received recommendations for further treatment from the consultant physician.’ means that there are no recommendations in the document.

Methods for recognising tokens and morphological annotation necessary to create the next two files *xxx_segm.xml* and *xxx_morph.xml* are described in the previous sections of this chapter, while semantic annotation methods are discussed in the next chapter.

Figure 5.5 gives a screen of the corpus editor that visualises data and allows for its modification. It contains a fragment of a diabetic document. At the bottom of the screen, the original text divided into paragraphs is given. The left part of the screen contains the sequence of tokens together with their interpretation. The editor allows the correcting of assigned morphological interpretations as well as introducing information on spelling errors (this part is not visible in the image). The right part of the screen contains the semantic annotation of data, which may also be modified. Moreover, we can see the fragment of text in which the semantic information is detected. It is highlighted in both text representations. The fragment describing the diabetes is highlighted in the text as at the right part of the screen we indicate an appropriate structure representing this information. A detailed explanation of the semantic annotation is given in the next chapter.

Tree Editor - d2006_072

File Edit Style

Ver.: 2.0

orth	base	ctag
53	53	number
.	.	interp
letni	letni	adj.sg.nom.m:1:pos
pacjent	pacjent	subst.sg.nom.m:1
z	z	prep:inst.nw:ok
cukrzyca	cukrzyca	subst.sg:inst:f
typu	typ	subst.sg.gen:m:3
2	2	number
.	.	interp
rozpoznaną	rozpoznao	ppas.sg:inst:f.perf.aff
przed	przed	prep:inst.nw:ok
10	10	number
.	.	interp
ma	ma	acron
laty	rok	subst.pl:inst:m:3
leczoną	leczyo	ppas.sg:acc:timperf.aff
insuliną	insulina	subst.sg:inst:f
i	i	conj
biguanidem	biguanid	subst.sg:inst:m:3
.	.	interp
z	z	prep:inst.nw:ok
nadciśnieniem	nadciśnienie	subst.sg:inst:n
tętnicznym	tętniczny	adj.sg:inst:n:pos
został	zostao	praet.so.m:1:perf

Discharge record

- EPIKRYZA_BEG
 - yes
 - ID_AGE
 - 53
 - feature_l_str
 - RELATIVE_DATA
 - D_TYPE
 - second
 - D_NUM
 - 10
 - D_UNIT
 - u_year
 - D_TREAT
 - insul_tr_t
 - ORAL_TREAT
 - biguanid_t
 - ACC_DISEASE
 - hypertension_t
 - ACC_DISEASE
 - hypertension_t
 - reason_l_str
 - D_CONTROLL
 - uncontrolled_t
 - D_CONTROLL
 - hiperglikemia_t
 - W_IN_WORDS
 - overweight
 - BMI
 - 36,5
 - HBA1C

146 Epikryza:
53- letni pacjent z cukrzyca typu 2, rozpoznaną przed 10-ma laty leczoną insuliną i biguanidem, z nadciśnieniem tętnicznym został przyjęty do Kliniki z powodu objawów niewyrównania cukrzycy - w postaci stałej hiperglikemii rzędu 200-300mg%, wzmożonego pragnienia i wielomoczu. Przy przyjęciu w stanie ogólnym dobrym. Z odchyłem od normy poza patologiczną otyłością BMI= 36,5 kg/m2 stwierdzono podwyższone ciśnienie tętnicze 180/100 mmHg, brak odruchów kołanowych i skokowych, zaburzenia czucia wibracji, temperatury i dotyku na stopach i w obrębie dolnej 1/3 części podudzi.

147

148 W badaniach biochemicznych poziom HbA1c= 11,4% potwierdza przewlekłe niewyrównanie cukrzycy.

149 W czasie pobytu zastosowano restrykcyjną dietę 1000-1200 kcal, zmodyfikowano dawkę i rodzaj insuliny uzyskując normalizację glikemii w dobowym profilu. Zredukowano dobową dawkę insuliny ze 120 j. na 88 j.dobę.

Ready...

Fig. 5.5: Corpus editor

Information extraction from hospital documents

In the chapter we discuss details related to the development of an information extraction system for medical texts in Polish. We present the system selecting information from diabetic patients' hospital records. The task is carried out with the help of manually created rules implemented in SProUT (Section 2.1.2). The system consists of a combination of: a sentence boundary detector, a tokeniser, and a morphological analyser. These components cooperate with a domain oriented gazetteer and manually prepared extraction rules.

Such systems extract predefined types of information usually identified on the basis of users' needs. It is necessary to determine types of information that should be selected, and values that might be assigned to them. For example, if we are looking for diabetic complications, we have to know a set of diseases that accompany the main illness and result from it. Apart from determining what kind of information we are looking for, it is necessary to determine relationships between particular pieces of information. All this data should be established in cooperation with experts. Extraction rules are defined based on expert knowledge and on an inspection of text. The system has to be evaluated on previously unseen data.

The results of the information extraction system can be used for annotation of hospital records with selected information. They can be applied to automatic construction of a semantically annotated corpus and to a swift search of it. Another application of the results of IE systems is filling of a database. Such a database can support the checking of the completeness of discharge documents, the preparing of statistics, the selection of cases that contain interesting data and the checking of dependences between occurrences of certain factors associated with the disease. Moreover, the results of the IE system can support preparation of data for machine learning experiments.

In the chapter we describe details related to developing the IE system. We start with a description of the information we are looking for and how it is represented as typed feature structures accepted by SProUT. Then, we describe the domain dictionary and reveal the extraction rules illustrating selected problems like recognition of several pieces of information in one context and recognition of negated information. After that, we show the evaluation of the method and the application of the results to the database creation and the annotation of texts. Finally, we describe the results of a machine learning experiment trained on data annotated with the help of our IE system.

6.1 Rule based information extraction

6.1.1 Domain description

For the purpose of information extraction from diabetic patients' hospital records, we elaborated a domain model that could represent data which we were interested in. The model was defined on the basis of an expert's knowledge and data, i.e. hospital documents. To formalise it we used the OWL-DL standard and the Protégé ontology editor (<http://protege.stanford.edu/>). In order to be used in the SProUT system the model was transformed into a typed feature structure (TFS) hierarchy that was required by the system.

To illustrate the complications of the extraction task, we quite precisely describe the data we are looking for below. The model represents the following information:

- Identification of a patient's visit to hospital, dates when it took place and reasons for hospitalisation. The reason for hospitalisation is only recognised when it is directly connected to the diabetes. The following explanation of hospitalisation: *przyjęty do kliniki z powodu wysokich poziomów glikemii* 'admitted to the hospital because of hyperglycemia' is such a reason. If a patient is admitted to hospital because of a broken leg it is not taken into account in the model.
- Patient's data: sex, age, height, weight (in numbers *75 kg* or words *otyły pacjent* 'obese patient') and BMI (Body Mass Index). Usually, this information is easy to identify. Sometimes it is a problem with gender identification, as the word *pacjent* may refer to a man or a woman. Normally, discharge records begin with the words *Pan* 'Mr.' or *Pani* 'Mrs.' which facilitates the task of gender recognition.
- Information concerning diabetes:
 - Diabetes type; we differentiate three types of the illness: first, second and other (if it is not one of the two main types of diabetes).
 - If the illness is balanced, e.g. *cukrzyca wyrównana* 'balanced diabetes' or *niewyrównana* 'unbalanced'.
 - When diabetes was first diagnosed. This information can be given in several ways: in words, e.g. *wieloletna* 'long-lasting'; as a date — *w 1990 roku* 'in the year 1990'; relatively *20 lat temu* '20 years ago'; or *w 20 roku życia* 'in the 20th year of life'. All these types of information demand different representation.
 - Results of basic tests, e.g. HbA1c, presence of acetone, levels of cholesterol, microalbuminuria and creatinine.
- Diabetes complications such as angiopathy with its micro and macroangiopathy subtypes, different types of retinopathy, types of neuropathy, nephropathy and diabetic foot; and other illnesses accompanying diabetes such as hypertension or autoimmunology illnesses.
- Diabetes treatment including:
 - Description of oral medications (we extract only their names).

- Information on the type of insulin (medication names) and doses. We also want to recognise information that insulin therapy has been started, which means that a patient is treated with insulin for the first time.
- Description of a continuous insulin infusion therapy.
- Diet description contains information on the type of diet and how many calories and meals are recommended. A typical phrase is: *Dieta cukrzycowa 2000 kcal, 5-6 posiłków/dobę* ‘Diabetics diet 2000 kcal, 5-6 meals/day’.
- Additional information important to the effectiveness of treatment, or some important changes. It is given in many different forms, so it is relatively difficult to identify. It concerns the following information:
 - Education: *Omówiono z chorą zasady diety, adaptacji dawek insuliny w zależności od różnych sytuacji* ‘Dietary rules, adaptation of insulin doses according to different situations was discussed with the patient’.
 - Diet observation: *Pacjent nie przestrzega diety* ‘The patient does not observe the diet’.
 - Self-monitoring: *Chory prowadzi samokontrolę poziomu glukozy we krwi.* ‘The patient self-monitors blood glucose levels’.
 - Therapy modification: *Zmieniono dotychczasowy system leczenia insuliny z dwóch wstrzyknięć na trzy* ‘The existing system of insulin therapy was changed from two to three injections’.

6.1.2 Information representation

The information is represented in SProUT by a multi-hierarchy of TFSs (typed feature structures). Below, we describe the type hierarchy designed for extracting information given in the previous section.

Let us consider structures defined to represent basic information concerning patient’s diabetes. For example, information when diabetes has been diagnosed is represented in the structure given in Figure 6.1. It is of the *diab_from_str* type and has four attributes that represent different ways of expressing time, all of them demanding a different coding.

$\left[\begin{array}{l} \textit{diabet_from_str} \\ \text{RELATIVE_DATA} \quad \textit{rel_data_str} \\ \text{ABSOLUTE_DATA} \quad \textit{data_str} \\ \text{YEAR_OF_LIFE} \quad \textit{string} \\ \text{FROM_IN_W} \quad \textit{d_from_word.t} \end{array} \right]$

Fig. 6.1: Structure of type *diab_from_str*

RELATIVE_DATA represents how many years, months or weeks ago diabetes has been diagnosed. This information is given, for example, by the following phrases: *20 lat temu* ‘20 years ago’, *przed 3 miesiącami* ‘for 3 months’. A

value of the attribute is another TFS of type *rel_data_str* with two attributes. The first one — `D_NUM` — is a number given as a string, and the second — `D_UNIT` — is the unit of time: year, month or week.

`ABSOLUT_DATA` represents the date when the illness was diagnosed. The value is a structure *dat_str* coding a day, month and year. Usually, this information is represented only by a year: *w 1990 roku* ‘in the year 1990’. But sometimes the date is given more specifically, e.g. *w czerwcu 1995* ‘in June 1995’ or as ‘05.12.1999’.

`YEAR_OF_LIFE` represents how old the patient was when diabetes was diagnosed, e.g. *w 20 roku życia* ‘in the 20th year of life’. Its value is represented as a string.

`FROM_IN_W` represents indirect information given in words, e.g. *wieloletnia* ‘long-lasting’, *świeżo rozpoznana* ‘newly recognised’. A value of the attribute is of the *d_from_word_t* type which is the supertype for three types: newly diagnosed, previously recognised and long-lasting diabetes.

Information about the diabetes type is represented by simple structures, given in Figure 6.2. We differentiate two main types of the disease: the *first* and *second* type. The specificity of other types is not distinguished and all of them, e.g. a gestational diabetes, are classified as the *other* type.

$$\left[\begin{array}{l} d_type_str \\ D_TYPE \quad d_type_t \end{array} \right]$$

Fig. 6.2: Simple structure of type *d_type_str*

Diabetic complications are also represented by a simple structure similar to the one in Figure 6.2 with one attribute `COMP` and the value of the type *complication_t*. The hierarchy in Figure 6.3 represents relationships among complications recognised by the grammar. Such a hierarchy allows the unification operation available in SProUT to be applied, see Section 6.1.4. For example, it is possible to refer in a rule to all subtypes of a complication using their supertype, i.e. if we refer to *microangiopathy_t* we also refer to its seven subtypes.

The hierarchy contains three main types of complications: neuropathy, angiopathy, diabetic foot and, additionally, the *no_complication* type. The first two types have subtypes. Neuropathy divides into autonomic neuropathy or peripheral polineuropathy, and angiopathy can be micro and macroangiopathy. Micro and macroangiopathy have further subtypes. Information about a common microangiopathy complication — retinopathy is represented by nonproliferative, preproliferative or proliferative retinopathy. Another microangiopathy complication is maculopathy.

Sometimes, it is necessary to recognise more than one piece of information through one rule. Then, the results may be represented as a list. We define three list structures in the diabetes type hierarchy. One is used for the representation of

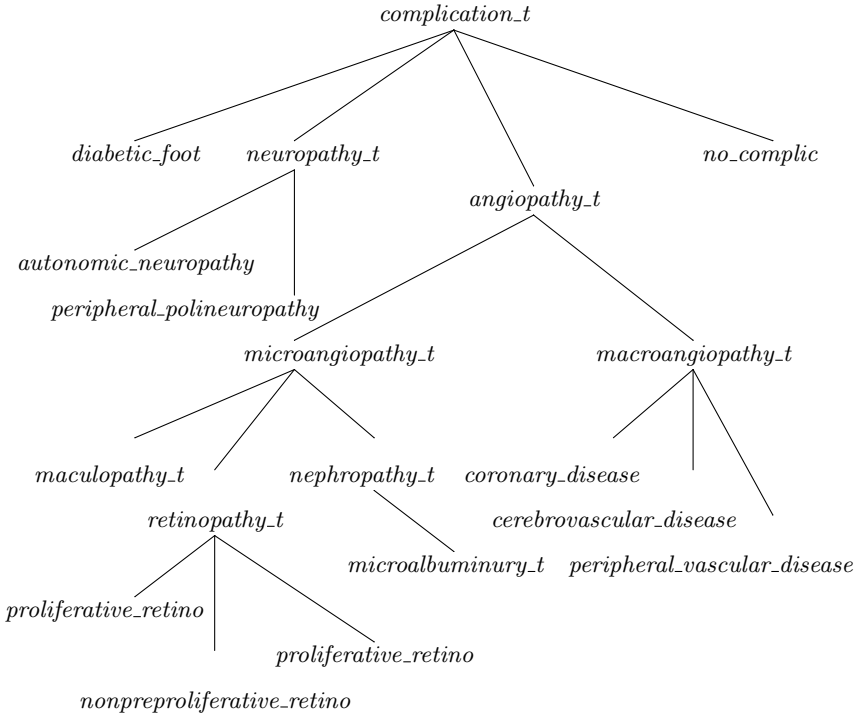


Fig. 6.3: Hierarchy of complications

different features of the disease (see Figure 6.4), the second, for the representation of reasons for hospitalisation, whereas the third, for the treatment description.

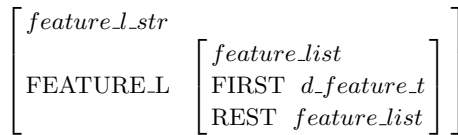


Fig. 6.4: List of features

As complications are subtypes of the feature type (*d_feature_t*), the list structure in Figure 6.4 represents, among other things, information contained in the following phrase: *z neuropatią autonomiczną i obwodową* ‘with autonomic and peripheral neuropathy’, see Figure 6.5.

Several features may have a dual role. For example, the phrase *cukrzyca źle kontrolowana* ‘poorly balanced diabetes’ is a statement of the diabetes feature, while the phrase *przyjęty do szpitala z powodu źle kontrolowanej cukrzycy* ‘admitted to the hospital because of poorly balanced diabetes’ also indicates the reason

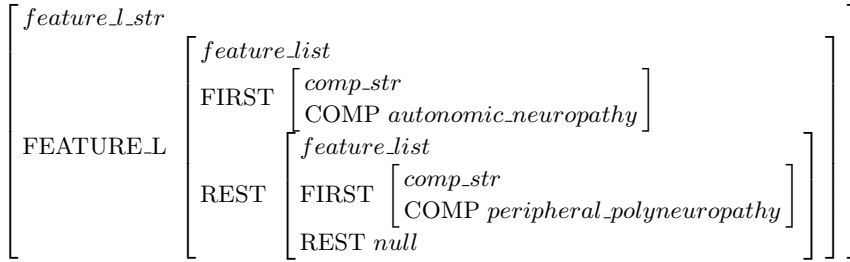


Fig. 6.5: Representation of ‘with autonomic and peripheral neuropathy’

for hospitalisation. We use a multi-hierarchy for coding such dual features. The type of diabetes balance information structure (d_contr_str) is defined simultaneously as the subtypes of the diabetes feature and reason for hospitalisation type.

6.1.3 Domain dictionary

In SProUT, it is possible to define a specialised lexicon called gazetteer. It contains important terms associated with semantic descriptions, e.g. concepts from the domain model. Such lexicons list terms not represented in general language dictionaries such as: medication names, diseases, examinations, abbreviations. Moreover, they list key concepts — words from a general dictionary but crucial in IE rules construction.

The fragment of a dictionary in Figure 6.6 contains insulin and oral medication names important in diabetes treatment as well as diseases and diabetic complications. Each line contains a form of a term and two levels of semantic information. Each form of a term is a string which may contain spaces and is ended with a separator. Thus, it is possible to define multi-word terms. The semantic information is represented by two attributes with values from the type hierarchy. The first one, *GTYPE* attribute, allows us to create groups of concepts useful in identifying pieces of information of the same type. *G_CONCEPT* attribute determines the type associated with the term. Various items of information in a line are separated by the ‘|’ symbol. As already mentioned, the dictionary contains forms of terms, so to limit its size, we add only forms occurring in the data. For example, as medications only appear in the text in the nominative and genitive case, only these two forms are included. The values of the *GTYPE* attribute in this fragment refer to the following types of information:

- *gaz_oral*: oral medication names important in diabetology;
- *gaz_insulin*: insulin medications;
- *gaz_comp*: diabetic complications;
- *gaz_neuro*: different types of neuropathy (a complication);
- *gaz_retino*: subtypes of retinopathy.

```

Monotard | GTYPE: gaz_oral | G_CONCEPT: monotard_t
Monotardu | GTYPE: gaz_oral | G_CONCEPT: monotard_t
Insulatard HM | GTYPE: gaz_insulin | G_CONCEPT:insulatard_hm_t
Insulatardu HM | GTYPE: gaz_insulin | G_CONCEPT:insulatard_hm_t
Novorapid | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
Novorapidu | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
NovoRapid | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
NovoRapidu | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
Novo Rapid | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
Novo Rapidu | GTYPE: gaz_insulin | G_CONCEPT:novorapid_t
neuropatie | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
neuropatię | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
Neuropatie | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
Neuropatię | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
Neuropatia | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
neuropatią | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
neuropatię | GTYPE: gaz_comp | G_CONCEPT: neuropathy_t
obwodową | GTYPE: gaz_neuro | G_CONCEPT: peripheral_polyneuropathy
obwodowa | GTYPE: gaz_neuro | G_CONCEPT: peripheral_polyneuropathy
obwodowej | GTYPE: gaz_neuro | G_CONCEPT: peripheral_polyneuropathy
autonomiczną | GTYPE: gaz_neuro | G_CONCEPT: autonomic_neuropathy
autonomiczna | GTYPE: gaz_neuro | G_CONCEPT: autonomic_neuropathy
autonomicznej | GTYPE: gaz_neuro | G_CONCEPT: autonomic_neuropathy
retinopatii | GTYPE: gaz_comp | G_CONCEPT:retinopathy_t
retinopatią | GTYPE: gaz_comp | G_CONCEPT:retinopathy_t
retinopatię | GTYPE: gaz_comp | G_CONCEPT:retinopathy_t
retinopatia | GTYPE: gaz_comp | G_CONCEPT:retinopathy_t
prosta | GTYPE: gaz_retino | G_CONCEPT:nonproliferative_retino
proliferacyjnej | GTYPE: gaz_retino | G_CONCEPT:proliferative_retino
proliferacyjną | GTYPE: gaz_retino | G_CONCEPT:proliferative_retino
proliferacyjna | GTYPE: gaz_retino | G_CONCEPT:proliferative_retino
proliferativa | GTYPE: gaz_retino | G_CONCEPT:proliferative_retino
przedproliferacyjnej | GTYPE: gaz_retino | G_CONCEPT:preproliferative_retino

```

Fig. 6.6: Fragment of a gazetteer

In a gazetteer, we can represent different forms and different spellings of terms to link them with the same concept. The following forms: *Monotard* (nominative) and *Monotardu* (genitive) are associated with the *monotard_t* type that is a subtype of the oral medication. The type *novorapid_t* is associated with six different forms that represent various spellings of this name with and without a space inside the name, and with and without the capital letter ‘R’ inside the strings.

The dictionary for the system extracting information from the diabetic patients’ hospital records contains 569 entries, including 274 forms of medication names, and 85 diabetic complications.

6.1.4 Grammar rules

An information extraction grammar consists of rules which are regular expressions over TFSs with functional operators and coreferences, representing recognition patterns.

Example of a rule

An example of a rule is given in Figure 6.7. The rule named *insulina* recognises the type and doses of an insulin treatment. Its result is a TFS with three attributes given in Figure 6.8. The value of the `L_TYPE` attribute is a subtype of the *gaz_insulina* type. The next two attributes are minimal and maximal doses (`DOSE_MIN` and `DOSE_MAX`), so these are numbers represented by the *string* type.

```

1: insulina :>
2:   @seek(liczba) & [LICZ #jedn1]
3:   (token & [TYPE hyphen] @seek(liczba) & [LICZ #jedn2])?
4:   token & [SURFACE "j"]
5:   token & [SURFACE "."]?
6:   gazetteer & [GTYPE gaz_insulin, G_CONCEPT #rodzaj]
7: -> insulin_treat_str & [L_TYPE #rodzaj,
8:   DOSE_MIN #jedn1,
9:   DOSE_MAX #jedn2].

```

Fig. 6.7: The rule recognising insulin with doses

<i>insulin_treat_str</i> <code>L_TYPE</code> <i>gaz_insulin</i> <code>DOSE_MIN</code> <i>string</i> <code>DOSE_MAX</code> <i>string</i>
--

Fig. 6.8: The structure representing insulin with doses

Line 1 contains the rule name. The next line contains the expression that recognises a number by a call (`@seek`) to another rule *liczba*. Its result is represented as the variable (`#jedn1`), which is unified with the value of the minimal dose attribute (`DOSE_MIN`) in the output structure. Line 3 is optional, as there is the ‘?’ symbol after the description of elements. It specifies two tokens — the first has the type ‘hyphen’ (it refers to the symbol ‘-’), while the second is a number. If the information described in line 3 is present, the value of the attribute `DOSE_MAX` is established and unified with the value of the `#jedn2` variable. Line 4 recognises the abbreviation of the unit, i.e. the ‘j’ letter, after which a dot may appear (the optional line 5). Finally, the rule refers to the gazetteer

and checks if the subsequent token is of the type *gaz_insuline*. The exact type of insulin (*#rodzaj*) is unified with the value of the attribute *L_TYPE* in the output structure.

(6.1) *Insulina:*

R 12 j Actrapid HM + 14 j. Insulatard HM

P 9 j. Actrapid HM

W 8 - 10 j. Insulatard HM

Furosemid 1 x 1 tabl.

Insulin:

Mo(rning) **12 j Actrapid HM + 14 j. Insulatard HM**

Mi(dday) **9 j. Actrapid HM**

E(vening) **8 - 10 j. Insulatard HM**

Furosemid 1 x 1 pill(ow).

Four triggers of the rule *insulina* recognise the four boldfaced excerpts (containing the dose and the insulin name) in the Example 6.1. In the case of the first three insulins, only the value of the *DOSE_MIN* attribute is established and the value of *DOSE_MAX* is left as *string* type. In the last case, the scope of the dose is defined and both attributes *DOSE_MIN* and *DOSE_MAX* have specified values.

Information recognised without context

Some information can be extracted only when we find a phrase expressing the fact. For example, a patient's weight is extracted without any additional conditions by the rule in Figure 6.9. If we find the key phrase *masa ciała* 'mass of body' or *waga* 'weight', the following number is interpreted as the patient's weight. The key phrases are recognised by the first line of the rule that refers to the morphological dictionary and the base forms of words (*STEM* attribute). The number is recognised by one of two rules: for natural numbers (*liczba_nat*), or decimal numbers (*liczba_ulam*) recognition. The attribute *WEIGHT* in the output structure is unified with the recognised number.

```
waga:>
  (( morph & [STEM "masa"] morph & [STEM "ciało"])    ;; 'mass of body'
   | morph & [STEM "waga"])                            ;; 'weight'
  (@seek(liczba_nat) & [LICZ #h] | @seek(liczba_ulam) & [LICZ #h] )
->weight_str & [WEIGHT #h].
```

Fig. 6.9: Weight recognition rule

Another example of information that doesn't require context is identification of different types of complications. The rule in Figure 6.10 identifies phrases related to neuropathy — a diabetic complication. It recognises both:

- occurrences of a single word *neuropatia* ‘neuropathy’ in all forms included in the lexicon;
- occurrences of a multi-word term that consists of a form of the word *neuropatia* ‘neuropathy’ and a word describing its subtypes *obwodowa* ‘peripheral’ or *autonomiczna* ‘autonomic’.

neuropatia:>

```
gazetteer & [GTYPE gaz_comp, G_CONCEPT neuropathy_t & #rodzaj]
(gazetteer & [GTYPE gaz_neuro, G_CONCEPT #rodzaj])?
-> comp_str & [COMP #rodzaj].
```

Fig. 6.10: Neuropathy recognition rule

The rule uses only information originating from the gazetteer, see Figure 6.6. The output structure has the attribute `COMP` whose value is unified with the value of the attribute `G_CONCEPT` of the first recognised element and the second one, if it is present. Unification guarantees that if both pieces of information are identified the more specific type is chosen.

Context information recognition

Information extraction usually can’t only be based on recognition of key words or phrases. The same word can be interpreted differently depending on the context. So, if we want to extract information as to whether diabetes is balanced, it is not enough to recognise the key word *niekontrolowana* ‘uncontrolled’. This word should be identified as representing this concept only in the context of the word *cukrzyca* ‘diabetes’. A similar problem arises when we want to identify when diabetes has been diagnosed, what type of diabetes it is, and what kind of treatment has been used. All this information requires the context of the *cukrzyca* ‘diabetes’ word. Otherwise, the following phrases could be misinterpreted: *niekontrolowane nadciśnienie* ‘uncontrolled hypertension’ or *wieloletnia nadczynność tarczycy* ‘long-lasting overactive thyroid’. Let us consider the following example:

(6.2) *Wieloletnia, źle kontrolowana, z retinopatią, cukrzyca typu 2*
 long-lasting, bad controlled, with retinopathy, diabetes type 2

Each word of this phrase carries important information: *wieloletnia* ‘long-lasting’, *źle kontrolowany* ‘badly controlled’, *z retinopatią* ‘with retinopathy’, *typ 2* ‘type 2’. All of them, except information about the retinopathy complication, should be identified as important only in the context of the key word *cukrzyca* ‘diabetes’. While there is only one word *cukrzyca* ‘diabetes’ in the phrase, all information should be recognised by one rule.

To account for this observation, we distinguish two groups of features according to the requirement of the word *cukrzyca* ‘diabetes’ in their context. The

first group consists of features that demand this word in the context. These are the following features together with examples of phrases representing them.

- Type of diabetes: *cukrzyca typu 1* ‘diabetes of type 1’ or *cukrzyca ciężarnych* ‘diabetes of pregnant’;
- When the diabetes has been diagnosed: *cukrzyca wykryta 5 lat temu* ‘diabetes diagnosed 5 years ago’, *świeżo rozpoznana cukrzyca* ‘newly recognised diabetes’;
- Type of treatment: *cukrzyca leczona dietą* ‘diabetes cured with a diet’;
- Balance of diabetes: *cukrzyca niekontrolowana* ‘unbalanced diabetes’, *cukrzyca źle wyrównana* ‘bad balanced diabetes’, *cukrzyca chwiejna* ‘unstable diabetes’.

```
cecha_terap:/
  (morph & [STEM "leczyć"]      ;; 'to cure'
  (token)?
  gazetteer & [GTYPE gaz_treat, G_CONCEPT #typ])
->feature_str & [FEATURE d_treat_str & [D_TREAT #typ]].
```

Fig. 6.11: Type of treatment recognition rule

Figure 6.11 contains the rule recognising a type of treatment which is a context depending feature. In the first line, the name of the rule *cecha_terap* is followed by ‘:.’. It means that the result of the rule is not directly outputted. In the fourth line, a gazetteer entry that indicates one of three types of treatment: diet, insulin or oral medications is recognised. In the diabetes grammar, the rule *cecha_terap* is called by another rule producing output with information about a type of treatment. It recognises the keyword *cukrzyca* ‘diabetes’, among others, to be sure that the treatment concerns this disease.

The second group consists of the following features that do not require the *cukrzyca* ‘diabetes’ context word. They can be recognised separately.

- Weight of a patient expressed in words: *otyły* ‘obese’, *z otyłością* ‘with obesity’ or just opposite *z niedowagą* ‘with deficit of weight’;
- Some phrases expressing problems with diabetes balance like *hipoglikemia* or *stany niedocukrzenia* ‘hypoglikemia’, *wahania glikemii* ‘shuffle of glycemia’;
- Information about *ketonuria* or *acetonuria*.

To recognise that a patient is obese, it is enough to find the phrase: *otyły* ‘obese’ or *z otyłością* ‘with obesity’ in the data. The rule in Figure 6.12 recognises words expressing an overweight or deficit of weight. All terms expressing this feature are inserted in the domain dictionary with the GTYPE of the *gaz_weight* value and the information *overweight* or *underweight* as a G_CONCEPT value

```

cecha_waga:>
(gazetteer & [GTYPE gaz_weight, G_CONCEPT #typ])
->feature_str & [FEATURE w_in_words_str & [W_IN_WORDS #typ]].

```

Fig. 6.12: Weigh expressions recognition rule

assigned to the *#typ* variable. This feature is recognised without any context, so the output is generated directly by this rule (‘:>’ after the name).

The solution for capturing all the information contained in Example 6.2 is to recognise the whole phrase through one rule which takes into account all 5 features together in the context of the key word ‘diabetes’.

Example 6.2 discussed in this section illustrates another problem — free word order in Polish. So, to express all the following phrases, we have to allow for all orders of the information. Our approach is to prepare rules covering all permutations of the information that is analysed together.

- *Wieloletnia, niekontrolowana cukrzyca typu 2,*
long-lasting uncontrolled diabetes type 2,
- *Niekontrolowana, wieloletnia cukrzyca typu 2,*
- *Wieloletnia cukrzyca typu 2, niekontrolowana,*
- *Cukrzyca wieloletnia typu 2, niekontrolowana.*

Negation

Hospital records usually contain information determining the existence of an illness, its complications or symptoms. Sometimes, equally important is information about a lack of them, as a lack of a complication may indicate that diabetes is not very advanced or the treatment is effective. The phrase *bez retinopatii* ‘without retinopathy’ isn’t correctly interpreted on the basis of the key word *retinopatii* ‘retinopathy’, we have to recognise negation expressed in the form of the negative preposition *bez* ‘without’.

The rule in Figure 6.13 captures a longer excerpt of text containing a complication and an expression that it is not diagnosed. For the phrase considered above, the output of the rule explicitly states that retinopathy is absent. In this rule, the important words that must appear are given in the first line where negation is expressed and in the last line where the type of complication is given. The second line is optional.

The negation may be expressed in more complex ways as in the following phrases that all mean that microangiopathy is not diagnosed:

- *nie występują późne powikłania cukrzycowe o charakterze mikroangiopatii* ‘there were no long-lasting diabetes complications of microangiopathy type’;
- *nie wykryto obecności późnych powikłań cukrzycowych pod postacią mikroangiopatii* ‘long-lasting diabetes complications in the form of microangiopathy were not detected’;

```

bez_zech :>
  (morph & [STEM "bez"] | morph & [STEM "brak"])      ;; 'without'
  ( morph & [STEM "postać"] | morph & [STEM "cecha"])?  ;; 'feature'
  gazetteer & [GTYPE gaz_comp, G_CONCEPT #t]
-> no_comp_str & [N_COMP #t].

```

Fig. 6.13: Rule recognising negation

- *nie stwierdzono późnych zmian cukrzycowych w postaci mikroangiopatii* ‘long-lasting diabetes complications in the form of microangiopathy were not found’
- *bez zmian o cechach mikroangiopatii* ‘without symptoms of microangiopathy type’.

To recognise them, it is necessary to create several rules that take into account longer phrases expressing negation.

Unification

Unification in a SProUT grammar is used almost in each rule to pass results to output structures. In Figure 6.13, the attribute value in the output structure is denoted by the variable ‘#t’. It is unified with a type of complication recognised with the help of the gazetteer (G_CONCEPT #t). Unification might also be helpful if we want to check an agreement between a noun and an adjective. However, we can often neglect it — in the whole diabetes grammar we don’t check agreement between phrase elements as we don’t check their correctness. For these phrases, the sequences of base forms have a unique interpretation. But it is not a general rule. The following phrases cannot be interpreted on the basis of the base forms only, and agreement is crucial.

- *tłumaczenie_{nom,neut} nowe_{nom,neut} książki_{gen,fem}*
‘a new translation of the book’
- *tłumaczenie_{nom,neut} nowej_{gen,neut} książki_{gen,fem}*
‘a translation of the new book’

As in medical phrases, this phenomenon is rather exceptional; we use only base forms (STEM) for phrase recognition in the grammar.

Sometimes, it is convenient to specify more complex type relations in order to take advantage of unification available in SProUT. Figure 6.14 gives an example of multi-hierarchy. It describes diet types, where a diet type may be a subtype of two other diet types. For example, a low fat diabetes diet is a subtype of a low fat diet and a diabetes diet. A rule that gains from this multi-hierarchy is given in Figure 6.15. It recognises the word *dieta* ‘diet’ (in any form: *diety_{gen}*, *dietę_{acc}*, *diacie_{loc}*) and up to three following tokens. Let us consider the phrase *dieta cukrzycowa niskotłuszczowa* (‘diet’ ‘diabetes’ ‘low-fat’) ‘low-fat diabetes diet’. The first token after the word *dieta* ‘diet’ is described as a diet type in the

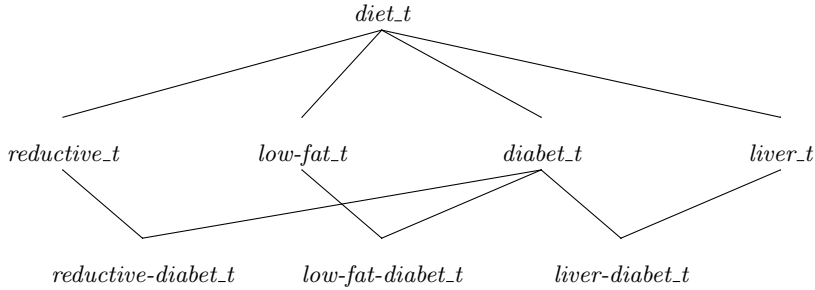


Fig. 6.14: Hierarchy of diet types

```

dieta :>
  morph & [STEM "dieta"]      ;; 'diet'
  gazetteer & [GTYPE gaz_diet, G_CONCEPT #d1]
  ((morph & [POS interp])?
    gazetteer & [GTYPE gaz_diet, G_CONCEPT #d1])?
-> diet_str & [DIET_TYPE #d1].

```

Fig. 6.15: Unification of diet types

gazetteer, e.g. *diabet.t* and is unified with the result — #d1 variable. The next token is optional and it is not included in recognition of the considered phrase. The last token analysed by the rule is also a diet type, e.g. *low-fat.t* and it is unified with the same #d1 variable. The result of unification of these two types is *low-fat-diabet.t* as expected.

6.1.5 Evaluation

To check the effectiveness of our extraction grammar we analysed 100 documents from data not inspected until the evaluation phase. They contained information appropriate for 55 attributes from the set of all 68 attributes, so we could only evaluate a subset of attributes. The evaluation, given in Table 6.1, was published in Mykowiecka *et al.* (2009). The overall results are good, the precision of all attributes is 99.26%, the recall — 96.49% and the F-score 97.86.

Identification of some complications is very good, for example all 72 cases of the two types of autonomic and peripheral neuropathy are fully recognised. A lower precision (94.21%) in recognising retinopathy is caused by negative expressions unforeseen by our grammar, such as *Przy nieobecności zmian o typie retinopatii* ‘With lack of changes of retinopathy type’. The attribute COMP (i.e. complication) occurs 369 times and has the F-score 98.39.

The information ‘when diagnosed’ coded by the *diab_from_str* structure (see Figure 6.1) represents when diabetes is first diagnosed. It can be expressed in many ways and needs the co-occurring word *cukrzyca* ‘diabetes’, so evaluation of its recognition is interesting. Sometimes the word ‘diabetes’ is omitted in

a sentence or is placed far from the fragment describing when diabetes was first diagnosed, which decreases recall. Some phrases are not recognised because of atypical abbreviations, perfectly understandable in the context by a human reader but not by a machine, since they are not present in our domain lexicons. For instance the phrase *cukrzyca rozpoznana w 25 r.ż.* ‘diabetes recognised in the 25th year of life’ is not recognised as *ż.* is an atypical, unpredicted, abbreviation for *życie* ‘life’.

The attribute D_CONTROL indicates if a therapy is effective. This information can be expressed by various phrases such as *niezrównoważona cukrzyca* ‘unbalanced diabetes’, *źle kontrolowana cukrzyca* ‘poorly controlled diabetes’, *hipoglikemia* ‘hypoglycemia’, *nieprawidłowy poziom glukozy* ‘abnormal levels of glucose’ or *cukrzyca dobrze kontrolowana* ‘well controlled diabetes’. As it is difficult to predict all such expressions, not all of them are recognised by the grammar rules. Recognition of the D_CONTROL attribute is sometimes correlated with the attribute REASON (for hospitalisation). Recall of REASON is lower (73.44%) due to long sentences which contain reasons mixed with other information. So, the same long phrases with mixed information also lower the recall of the attribute D_CONTROL.

The attribute RECOMMENDED_DIET consists of two pieces of information: how many calories and how many meals are recommended. Both values can be expressed as ranges. The calories are usually recognised correctly but typographical problems (non-standard abbreviations or a lack of space between the number and the word *posilek* ‘meal’) lowered the recall to 85.71%.

Some attributes are rare in the test set (e.g. SELF_MONITORING, DIET_CORRECTION, DIET_OBSERVATION), so interpretation of their recognition results is not reliable. There are 14 rare attributes that occur less than 10 times in the data. For 9 of them we have F-score equal to 100 — all of them are correctly recognised, but the single occurrences of 2 attributes were not recognised which results in 0% recall and precision.

Table 6.1: Evaluation results for attributes found in 100 test documents

	Attribute	Cases	Prec.	Recall	F-score
ALL ATTRIBUTES		4,021	99.26	96.49	97.86
Document data					
begin	DOC_BEG	100	100	100	100
date	DOC_DAT	100	100	100	100
identification	ID	100	100	98	98.99
continuation	CONT	98	100	97.96	98.97
hospitalisation from	H_FROM	99	100	98.99	99.49
hospitalisation to	H_TO	99	100	98.99	99.49
beginning of summary	EPICRISIS	100	100	100	100
Patient data					
identification	ID_PAT	100	99.01	100	99.50
sex	ID_P_SEX	101	100	100	100
age	ID_AGE	192	100	98.44	99.21

Table 6.1: (continued)

	Attribute	Cases	Prec.	Recall	F-score
height	HEIGHT	87	100	98.85	99.42
weight	WEIGHT	88	100	100	100
BMI	BMI	88	98.82	95.45	97.11
weight in words	W_IN_WORDS	62	98.41	100	99.20
Diabetes features					
type	D_TYPE	173	98.84	98.27	98.55
balance	D_CONTROL	212	98.54	95.75	97.13
acetonuria	ACET	11	91.67	100	95.65
treating method	D_TREAT	50	100	92	95.83
when diagnosed					
year	Y_DAT	6	100	100	100
relative data	D_NUM	41	100	97.56	98.77
	D_UNIT	41	100	97.56	98.77
year of life	YEAR_OF_LIFE	3	100	100	100
n words	FROM_IN_W	32	93.94	96.88	95.38
Diseases					
all diab. complications	COMP	369	97.35	99.46	98.39
retinopathy type	RETINOPATHY_T	120	98.36	100	99.17
maculopathy type	MACULOPATHY_T	32	100	90.63	95.08
no compl. of type	N_COMP	34	100	100	100
accompanying	ACC_DISEASE	141	100	100	100
autoimmune	AUTOIMM_DISEASE	4	100	100	100
Test results					
creatinine lev. 1	CREATIN1	96	100	100	100
creatinine lev. 2	CREATIN2	3	60	100	75
HbA1C	HBA1C	146	100	93.15	96.45
cholesterol LDL	LDL	81	100	100	100
microalbuminury	LEV1		100	92.59	96.15
Diet recommended					
min. calories	CAL_MIN	102	100	94.12	96.97
max. calories	CAL_MAX	4	100	50	66.67
min meals	MEALS_MIN	95	100	87.37	93.26
max meals	MEALS_MAX	20	100	80	88.89
Insulin therapy					
dose description					
insulin medication	I_TYPE	298	100	92.62	96.17
min. dose	DOSE_MIN	298	100	90.60	95.07
max. dose	DOSE_MAX	35	100	97.14	98.55
continuous infusion					
insulin medication	INS_TYPE	2	100	100	100
min. basal per day	TOT_MIN_BASE	1	100	100	100
max. basal per day	TOT_MAX_BASE	2	100	100	100

Table 6.1: (continued)

	Attribute	Cases	Prec.	Recall	F-score
min. bolus per meal	B_MIN	2	100	100	100
max. bolus per meal	B_MAX	2	100	100	100
Oral medication	ORAL_TREAT	76	96.15	98.68	97.40
Various					
reason for hospit.	REASON	83	98.73	93.98	96.30
training of patient	EDUCATION	46	100	91.30	95.45
insulin therapy beg.	THERAPY_BEG	3	66.67	66.67	66.67
therapy modification	THERAPY_ MODIF	23	100	91.30	95.45
insulin dose modif.	DOSE_MODIF	10	90	90	90
self monitoring	SELF_MONITORING	1	0	0	-
diet correction	DIET_CORRECTION	1	100	100	100
diet observing	DIET_OBSERVE	1	0	0	-

6.2 Database

Results of an information extraction system can be used to automatically fill a database. It allows for quick searches for a specific disease; patients having a certain test result; preparation of statistics; or testing hypothesis. The results of the diabetic information extraction system have been used to prepare a relation database described in (Marciniak *et al.*, 2008). In this section, we draw attention to some issues that should be taken into account while inserting extracted information into a database, as the IE results cannot be directly used as an input file for a program filling a database. For example, we need to remove redundant or non-informative pieces of extracted structures and to normalise data that is recognised in different formats.

An information extraction system in SProUT produces results as an XML file which contains, among important information, structures having the attribute value of the most general type that do not bring any information. Some information is repeated several times in text, so redundant output structures should be removed. As the system allows for unification, some values are represented by labels that need to be resolved and we have to find attributes whose values are assigned to the labels. Figure 6.16 gives preliminary cleaned results of extracted information from a document.

The program filling the diabetic database also performs the following tasks:

- It checks if a document is complete. Information from documents containing only a part of obligatory data are not inserted into the database.
- Only documents of diabetic patients are inserted into the database, so it selects documents which contain information about diabetes type.
- Documents and their appendices are joined, as it may happen that a document containing some test results is created separately as a continuation of the main document.


```

DOC_BEG:yes
CONT:no||ID:16628||ID_YEAR:string
ID_AGE:67||ID_PAT:d2006_137||ID_P_SEX:masc.t
D_TYPE:second||UNCONTROLLED:yes
COMP:nephropathy.t
COMP:autonomic_neuropathy
COMP:peripheral_vascular_disease
ACC_DISEASE:hypertension.t
BMI:28,5||
ACETON_IN_URINE:no
CREATIN_1:1,6||CREATIN_2:string||CREATIN_3:string
DOSE_MAX:string||DOSE_MIN:16||I_TYPE:insulatard_hm.t
HBA1C:10,0
COMP:nonproliferative_retino
COMP:nonproliferative_retino||COMP:maculopathy.t
EPIKRYZA_BEG:yes
DIAB_FROM:diab_from_str||DIAB_FROM|DIAB_FROM|RELATIVE_DATA|D_NUM:23||
  DIAB_FROM|RELATIVE_DATA|D_UNIT:u_year||D_TYPE:second||
UNCONTROLLED:yes
COMP:nephropathy.t
COMP:nonproliferative_retino||COMP:maculopathy.t
COMP:autonomic_neuropathy
DOSE_MAX:string||DOSE_MIN:10||I_TYPE:actrapid.t
DOSE_MAX:string||DOSE_MIN:16||I_TYPE:insulatard_hm.t

```

Fig. 6.16: Preliminary cleaned results of extracted information

- Only important information is selected. For example, information about a diabetes treatment appears in a document several times, as all attempts to choose the treatment are recorded, but the final recommendation is given after the word *Epikryza* and only this treatment is inserted into the database. If a document contains information about the positive result of a test on acetonuria, it probably also contains information about the lack of it after treatment. Acetonuria indicates a serious medical condition that has to be eliminated before a patient release. We only insert information about acetonuria recognition into the database and not the lack of it.
- Normalisation of some data. For example, information when diabetes was first diagnosed can be given in many ways. So, information ‘20 years ago’ and when a patient was ‘20 years old’ has to be represented in absolute dates, as only such a form allows us to interpret information effectively.
- Introducing only the most specific information to the database. So, if a patient is diagnosed with retinopathy, the system infers that he/she also has a more general complication, e.g. angiopathy (see Figure 6.3).

The database allows us to give a quick answer to many important questions. Some examples are given below:

- If a hospital record contains all crucial data. For example, all records should contain information when a patient was diagnosed with diabetes.
- What percentage of patients with diabetes type 2 and with HbA1C greater than 7.5 are treated with insulin.
- How many patients with creatinine levels above 1.3 are suffering from microvascular (nephropathy, retinopathy) complications.

6.3 Automatic semantic annotation of corpora

Corpus annotations may be created manually from scratch or we can use an automatic annotation and manually correct them. Manual annotation of data is usually long and laborious and it is prone to errors. Manual annotation standards consist in preparing and accepting annotation guidelines; annotation of every text by at least two annotators; and resolving differences by a third experienced annotator. Manual annotation takes time and annotators may change their judgments in the course of work. Newly identified problems may influence points of view on the previously undertaken decisions or may require the extending of the annotation structure or guidelines. Moreover, manually constructed data is very hard to extend and modify — every change imposes extra effort for checking the consistency of all data. Rule based annotation needs preparation of an appropriate set of rules, but for a trained person it is faster and results are consistent. Although it does not guarantee complete correctness, the cost of correcting already labelled data is lower than the cost of entirely manual annotation.

The information extraction system described in Section 6.1 is not ready for direct application for corpus annotation. The results need further processing as the main goal of an information extraction task is to find out whether a particular piece of information is present in data, while a corpus annotation task requires identification of the boundaries of text fragments which are to be annotated with a given label. To solve the problem we combine the results of two extraction grammars. One extracts precise text pieces, while the second verifies their correctness by a more complex grammar which takes into account contexts of recognised information. Both grammars work on the original text and link text fragments to the extracted structures. This approach is similar to the idea of cascade grammars, where the next level operates on the results of a previous one.

6.3.1 Method

The task of automatic annotation of diabetic corpus with semantic information is carried out in several steps. First, the extraction grammar rules described in Section 6.1 are applied to documents. The rules recognise information with a high precision and recall, but if a complex structure consists of several attributes, they are all assigned to an entire phrase. We don't know which part of the phrase

represents a particular attribute. Let us return to Example 6.2 *Wieloletnia, źle kontrolowana, z retinopatią, cukrzyca typu 2* ‘Long-lasting, bad controlled, with retinopathy, diabetes type 2’. In Section 6.1.4 we argue for the need to identify the features of diabetes together by one complex rule. Now, we want to know which fragment of the phrase is associated with which diabetes feature.

To obtain more precise information, we apply simplified extraction grammar rules to the text. The rules recognise small pieces of information, usually one attribute (only numerical values are still recognised together (e.g. ranges or dates)). The grammar is developed by removing all context information from the original grammar rules. Thus, it recognises attributes with the same high recall, but the precision is rather low as it recognises too much information. The rules do not check any context, so for example the phrase *10 lat temu* ‘10 years ago’ is always recognised by this grammar as information when diabetes is diagnosed, as we assume that all such phrases concern diabetes. It can refer to another ailment like *nadciśnienie zdiagnozowane 10 lat temu* ‘hypertension diagnosed 10 years ago’ or other type of information such as *hospitalizowany 10 lat temu* ‘hospitalised 10 years ago’, see examples comparing results of both grammars in Appendix B.

The results of both extraction grammars are cleaned up in a way similar to that described in Section 6.2. Non-informative pieces of structures are removed from the results. For example, if in a diet description only one quantity of calories is given, it is assigned to the CAL_MIN attribute, while the CAL_MAX attribute is assigned a general *string* value that can be removed.

After cleaning, the results of both grammars are compared and only common structures are represented in the final data. We take into account the boundaries of phrases from both grammars. The broad phrases (from the full grammar) indicate the context in which the narrow phrases (from the simplified grammar) should be taken into account in the final annotation.

Finally, we correct some inconsistencies in recognised phrases. For example, the abbreviation of a year (*r*) is sometimes recognised with, and sometimes without, a subsequent dot. In the case of creatinine level test results, only a part of a complex unit is included in the phrase to which the output structure is attached. Such inconsistencies could cause problems in determining the principles for a manual verification of an annotation and should be removed.

The method results in annotation of 66,165 simple attribute values, while 43,789 of them are elements of complex feature-value structures. Table 6.2, cited earlier in (Mykowiecka and Marciniak, 2011a), shows how many occurrences of selected types of structures are identified in the corpus. Simple attributes that may occur both in isolation and inside complex structures are counted together.

Table 6.2: Semantic labels occurrences

Structure/Attribute	Nb. of occur.	Nb of labeled tok.	Nb of possible values	Nb of diff. phr. types
DOC_BEG	460	920	1	2
DOC_DAT	390	3,119	(date)	3
<i>id_str</i>	457	1,833	(number/symbol)	3

Table 6.2: (continued)

Structure/Attribute	Nb. of occur.	Nb of labeled tok.	Nb of possible values	Nb of diff. phr. types
<i>hospit_str</i>	436	7,821	–	11
H_FRON	436	3,053	(date)	3
H_TO	436	3,056	(date)	2
EPIKRYZA_BEG	459	459	1	1
<i>recommendation_str</i>	445	2,471	–	57
RECOMMENDATION_BEG	443	886	1	2
<i>id_pat_str</i>	443	1,838	–	3
ID_P_SEX	443	944	(symbol)	3
ID_AGE	903	2,237	(number)	6
W_IN_WORDS	247	247	2	8
WEIGHT	390	1,424	(number)	6
BMI	328	1,101	(number)	7
HEIGHT	390	807	(number)	5
D_TYPE	738	1,461	3	10
D_CONTROLL	867	1,615	4	81
ABSOLUT_DATA	11	33	(date)	3
FROM_IN_W	146	146	2	9
RELATIVE_DATA	153	463	(number/unit)	22
YEAR_OF_LIFE	12	49	(number)	4
HBA1C	511	2,519	(number)	12
KWAS_D	18	36	2	2
ACET_D	469	987	2	24
KETO_D	11	12	2	4
<i>creatinin_str</i>	430	1,865	(numbers)	11
<i>microalbuminury_str</i>	80	413	(numbers)	16
<i>lipid_str</i>	259	6,270	–	23
LDL1	247	540	(number)	3
<i>reason_l_str</i>	279	2,997	–	151
ACC_DISEASE	491	491	1	3
COMP	1,327	2,654	16	122
N_COMP	191	1,020	18	32
AUTOIMM_DISEASE	13	16	3	2
<i>insulin_treat_str</i>	4,387	24,896	–	321
L_TYPE	4,910	7,987	51	61
dose_str	4,339	13,330	(range)	16
<i>diet_str</i>	671	3,608	–	93
DIET_TYPE	421	857	7	9
cal_str	421	1,149	(range)	12
meals_str	388	890	(range)	9
<i>insulin_inf_treat</i>	31	220	(range)	14
<i>bolus_b_m</i>	7	40	(range)	7

Table 6.2: (continued)

Structure/Attribute	Nb. of occur.	Nb of labeled tok.	Nb of possible values	Nb of diff. phr. types
ORAL_TREAT	1,026	1,177	36	41
D_TREAT	275	581	3	28
L_THERAPY_BEG	27	88	1	11
THERAPY_MODIFF	184	711	2	90
DOSE_MODIFF	65	195	2	30
DIET_CORRECTION	18	51	2	12
DIET_OBSERVE	7	22	2	2
SELF_MONITORING	13	36	2	8
EDUCATION	355	2,914	1	189

6.3.2 Guidelines for manual verification

The guidelines for manual verification of the semantic annotation of the diabetic corpus describe all structures recognised by both grammars, with a precise definition of a phrase to which the structure might be assigned. Structures are assigned to continuous phrases, i.e. to all tokens between the first and the last token of the phrase. The precise definition of phrase boundaries consists in determining the sets of words that may start and end the phrase, and the type of information included in the phrase. This allows for the determination of all phrases that are taken into account by the grammars. Unfortunately, there are phrases that represent information that should be taken into account but are not predicted by the grammar designer. In such cases, annotators have to rely on their own opinion as to which words belong to such a phrase. According to these guidelines, annotators should correct all labels together with the boundaries of phrases linked to them.

Annotators have to point out not only information that is constructed according to the guidelines, but also other ways of representing adequate information if it is understandable to a human reader. For example, the following phrase: *Dieta cukrzycowa 6x (2100 kcal/dobę)* ‘Diabetic diet 6x (2100 kcal/24 hours a day)’ is easily interpreted by a person as 6 meals per day with a total quantity of 2100 kcal recommended, but is difficult to be predicted by a grammar designer. Another example illustrates a situation where knowledge about a document structure is necessary for the appropriate interpretation of a phrase. If the phrase *Bez późnych zmian cukrzycowych* ‘There are no long-lasting diabetes complications’ is included within an eye test description, it means that there are no complications diagnosed within the eyes. So, the information about the lack of complications should be limited to eye complications like retinopathy. On the other hand, if it is included in the final part of the document, it means that none of the long-lasting diabetic complications are diagnosed.

Another important guideline for annotators recommends ignoring understandable spelling errors. These errors may mean that information is not recognised by the grammars. For example, we recommend annotators to interpret

the word *pRetinopatia* ‘pRetinopathy’ as ‘Retinopathy’ and recognise it as the complication.

6.3.3 Evaluation

Table 6.3: Semantic label diversity and label verification

Structure/Attribute	Numb of occur.		F-score	Recall	Phr. types	Avg length	Words total
	Gold	rules					
DOC_BEG	46	46	100	100	1	2	92
DOC_DAT	37	38	99	100	1	1	298
id_str	46	45	99	98	2	4	183
ID	46	45	99	98	–	–	–
CONT	45	45	100	100	–	–	–
hospit_str	46	43	97	93	7	18	831
H_FROM	46	43	97	93	2	7	323
H_TO	46	43	97	93	2	7	323
EPIKRYZA_BEG	46	46	100	100	1	1	46
recommendation_str	44	44	100	100	13	5.6	248
RECOMMEND_BEG	44	44	100	100	1	2	88
id_pat_str	46	45	99	100	11	4	191
ID_PAT	46	45	99	100	–	–	–
ID_P_SEX	46	45	99	100	–	–	–
ID_AGE	46	45	99	100	6	2	91
W_IN_WORDS	6	5	91	83	4	1	6
WEIGHT	40	39	99	97	6	3.5	138
BMI	33	33	100	100	3	3.6	119
HEIGHT	40	39	99	97	3	2	80
D_CONTROLL	30	27	95	90	18	1.8	55
FROM_IN_W	1	0	–	–	1	2	2
HBA1C	59	54	96	92	8	5	299
ACET_D	42	42	100	100	4	2	85
creatinin_str	43	41	98	95	7	4.4	191
microalbuminury_str	13	12	96	92	6	5	65
lipid_str	31	27	93	87	6	27	834
LDL1	31	27	93	87	3	2.3	78
feature_l_str	91	91	100	100	59	5.7	518
COMP	5	5	100	100	4	1.6	8
D_CONTROLL	34	34	100	100	9	1.2	40
D_TREAT	24	24	100	100	6	2	49
D_TYPE	70	70	100	100	2	2	139
FROM_IN_W	10	10	100	100	5	1	10
RELATIVE_DATA	19	18	97	95	7	3	58
W_IN_WORDS	10	10	100	100	4		10
reason_l_str	30	27	95	90	27	12.3	370

Table 6.3: (continued)

Structure/Attribute	Numb of occur.		F-score	Recall	Phr. types	Avg length	Words total
	Gold	rules					
D_CONTROLL	40	37	85	95	19	2.3	94
KETO_D	2	2	100	100	2	1	2
KWAS_D	1	1	100	100	1	2	2
RELATIVE_DATA	1	1	100	100	1	4	4
SELF_MONITORING	1	1	100	100	1	1	4
ACC_DISEASE	48	48	100	100	3	1	48
COMP	134	132	97	96	49	2.2	294
N_COMP	27	15	71	56	11	5	134
insulin_treat_str	446	444	99	99	103	5.7	2,531
I_TYPE	439	436	99	99	23	1.7	746
dose_str	441	440	99	99	8	3.1	1,363
corr_str	2	1	67	50	2	6	12
DOSE_MODIFF	2	1	67	50	1	1	2
THERAPY_MODIFF	2	1	67	50	1	2.5	5
diet_str	47	44	97	94	29	7.8	366
DIET_TYPE	47	44	97	94	4	2.1	100
cal_str	47	44	97	94	6	2.8	131
CAL_MIN	47	44	97	94	–	–	–
meals_str	45	41	95	91	8	2.2	99
MEALS_MIN	45	41	95	91	–	–	–
ORAL_TREAT	63	63	100	100	18	1.2	75
I_THERAPY_BEG	4	1	40	25	4	5.3	21
THERAPY_MODIFF	24	19	88	79	23	4.3	103
DOSE_MODIFF	9	8	94	89	6	3.3	30
DIET_CORRECTION	2	2	100	100	2	3	6
SELF_MONITORING	0	2	–	–	3	1	3
EDUCATION	27	25	96	93	20	8	215

Manual verification of 10% of the corpus, a randomly selected 46 records consisting of 46,439 tokens, was done by two annotators. One coherent version was negotiated and accepted as a Gold-standard version. Kappa coefficient for annotators' agreement counted for all word-label pairs was equal to 0.976 if empty labels were counted and 0.966 when they were ignored (9,031 occurrences).

The number of differences between the automatic annotation and the Gold-standard concerned 596 tokens, i.e. 1.3% of labels. Kappa coefficient for the Gold-standard version and the automatically annotated set was equal to 0.94. Changes made by annotators mainly concerned:

- addition of new labels (for 79 structures, 554 tokens).
- deletion of mistakenly recognised structures (for 4 structures, 20 tokens);
- a few changes of the boundaries or the name of the structure.

The detailed results of evaluation, containing statistics of occurrences of attributes, the recall and F-scores values for all attributes and structures identified in the evaluation set, are given in Table 6.3, cited from (Mykowiecka and Marciniak, 2011b). The results of the comparison of the automatically annotated corpus and the manually corrected version counted on 3,309 structures (simple or complex) show 96.1% accuracy, precision – 99.4%, recall – 96.6%, F-measure – 98. The results of the verification counted on 9,057 non empty word–label pairs are as follows: accuracy – 98.7%, precision – 99.5%, recall – 93.6%, F-measure – 96.6.

Improperly recognised (not reflected in the data) information concerns only four types of structures, while the precision of all other attributes is equal to 100%, so if they are recognised, they are recognised correctly. An example of misinterpretation concerns the phrase *cukrzyca typu 2* ‘diabetes type 2’. IE grammar interprets it, without exceptions that a patient has diabetes type 2. But for the following phrase *pacjent obciążony rodzinie, mama i babcia z cukrzycą typu 2* ‘patient with family burden, mother and grandmother with diabetes type 2’, it is not true.

Below, we mention some reasons for errors resulting from unrecognised information. These errors influence the recall of evaluation.

- Unpredicted values of attributes, e.g. *dieta cukrzycowa wysokobiałkowa 188 kcal, 3 posiłki* ‘diabetic high protein diet 1800 kcal, 3 meals’ – we don’t recognise a diet of ‘diabetic and high protein’ type;
- Spelling errors or punctuation errors occurring in words crucial to firing rules:
 - *wlew podstawowy* instead of *podstawowy* ‘base infusion’;
 - *pRetinopathia* instead of *Retinopathia* ‘Retinopathy’;
 - *masa ciała103* ‘weight103’.
- Information represented by phrases not predicted by the extraction grammar. For example, we don’t label information on the obesity of a patient, when it is expressed in Latin ‘obesitas’ instead of Polish ‘otyłość’: *Warunki badania trudne obesitas* ‘Difficult obesity test conditions’.

The quality of the results obtained is good enough for using this data for a machine learning approach to the information extraction task that is described in the next section.

6.4 Machine learning experiment

Machine learning methods of information extraction are much easier to reuse in various domains and languages. But they need appropriate annotated data to learn from, the preparation of which is time-consuming.

As we had annotated diabetic corpus, we performed a machine learning experiment with the Conditional Random Fields method. It was successfully used within the LUNA project (Marciniak, 2010a) to annotate Polish dialogues from a public transport call centre with domain concepts, see (Mykowiecka and

Waszczuk, 2009), and (Waszczuk, 2010). In the experiment, we used the implementation prepared within the LUNA project, but there are many publicly available implementations of the method (see e.g.: CRF++ <http://taku910.github.io/crfpp/>; CRF Project Page implemented by Sunita Sarawagi <http://crf.sourceforge.net/>; CRFsuite (Okazaki, 2007)).

The entire experiment is presented in (Mykowiecka and Marciniak, 2011a). Here, we describe the manner of data preparation for the experiment and quote the best results obtained for the data.

6.4.1 Data

```

&      # &      #interp#
Po      # po     #prep:loc#
podaniu # podać    #ger:sg:loc:n:perf:aff#
furosemidu # Furosemid #subst:sg:gen:m3#
nastąpił # nastąpić #praet:sg:m3:perf#
niewielki # niewielki #adj:sg:nom:m3:pos#
przejściowy # przejściowy #adj:sg:nom:m3:pos#
wzrost    # wzrost   #subst:sg:nom:m3#
poszerzenia # poszerzenie #subst:sg:gen:n#
ukm       # układ moczowy #brev:npun:nphr#
nerki     # nerka     #subst:sg:gen:f#
prawej    # prawy     #adj:sg:gen:f:pos#
-         # -         #interp#
po        # po        #prep:loc#
10        # 10       #number#
min       # min      #unit#
od        # od       #prep:gen:nwok#
podania   # podać    #ger:sg:gen:n:perf:aff#
F         # F        #brev:pun:nw#
delikatne # delikatny #adj:sg:acc:n:pos#
zmniejszanie # zmniejszać #ger:sg:nom:n:imperf:aff#
się       # się      #qub#
poszerzenia # poszerzenie #subst:sg:gen:n#
.         # .        #interp#
&        # &       #interp#

```

Fig. 6.17: Fragment of a hash file

The structure of the corpus given in Chapter 5.9 is difficult to operate on, as information of different levels is contained in separate files. To overcome this problem, we create one, easy to process, file with all necessary information that

may be important in learning models. This file we call the “hash file” as we use the hash (‘#’) sign to separate information. The file contains three columns. The first one contains subsequent tokens. The second column contains the lemma of the token if it is a word, or the same token if it is a number or a punctuation mark. The end of a sentence is indicated by the ‘&’, mark as dots are ambiguous. They are used to separate the integer part from the fractional part of a number, or may appear within abbreviations. The last column contains a POS and morphological description if it is a word. Otherwise it contains a token type.

< <i>niedostateczne wyrównanie cukrzycy</i> > ‘uncontrolled diabetes’	$\left[\begin{array}{l} d_contr_str \\ D_CONTROLL\ uncontrolled_t \\ < \textit{niedostateczne wyrównanie} > \end{array} \right]$
--	---

Fig. 6.18: Simple structure with the attribute D_CONTROLL

< *przyjęty do kliniki z powodu niewyrównania cukrzycy* >
 ‘admitted to the hospital because of uncontrolled diabetes’

$\left[\begin{array}{l} reaso_l_str \\ \\ REASO_L \end{array} \right]$	$\left[\begin{array}{l} reaso_list \\ \\ FIRST \left[\begin{array}{l} d_contr_str \\ D_CONTROLL\ uncontrolled_t \\ < \textit{niewyrównania} > \end{array} \right] \\ \\ REST\ null \end{array} \right]$
--	--

Fig. 6.19: List structure with the attribute D_CONTROLL

As training data for this experiment, we use annotation obtained by methods described in Section 6.3. But, for CRF models, we have to simplify the labels of structures, as they can assign only one label to every token. Thus, we ‘linearise’ complex structures. We have two solutions we can use as labels: the most internal attributes or the entire paths. The SProUT system requires unique names of attributes in all structures, so the same labels may appear only if the output is a simple structure or a list of structures as in Figures 6.18–6.19. If we operate on simple attributes, for both pieces of texts: *niedostateczne wyrównanie* and *niewyrównanej* ‘unballanced’ we use the same D_CONTROLL label, even though in the first case it is general information, while in the second case it is also the reason for hospitalisation. If we use full path labels, in annotation of the second example the following label REASO_L|D_CONTROLL¹ is attached to the text fragment.

¹ The attributes indicating list elements (i.e. FIRST, REST) can be omitted.

The example of annotation of text with internal attributes looks as in Figure 6.20, where each token is annotated with one of the following tags: the beginning of the fragment annotated with a label, a continuation of a fragment annotated with a label, or the NULL tag indicating that no labels are assigned to the token. Please note that the first token *Cukrzyca* is annotated with the label indicating the beginning of a feature structure list.

Cukrzyca	#	<i>feature_L_str</i> -begin
typu	#	D_TYPE-begin
1	#	D_TYPE-continuation
długością	#	FROM_IN_W-begin
z	#	NULL
chwiejnością	#	D_CONTROLL-begin
.	#	NULL

Fig. 6.20: Semantic annotation for CRF

6.4.2 Label assignment with CRF model

The specificity of the CRF method allows us to build models on the basis of many features describing particular input items. As these features, we can use: token forms, POS names, lemmas, morphological features and token types. Features may concern an annotated element or its contexts.

The machine learning experiment requires separate data sets to learn, train and evaluate. So, we divided data consisting of 460 hospital documents into three sets: a training set consisting of 368 documents, a test set of 46 documents, and 46 documents as an evaluating set (manually corrected, see Section 6.3.3). We tested several combinations of up to 8 features per token. The best results we obtained for internal attribute labels (i.e. consisting of the most deeply nested attributes instead of the full path) taking into account the following features: a POS, a lemma and two preceding and two following tokens.

The results for the best model, sorted from the best to the worst F-score, are given in Table 6.4. They show nearly 94 of F-measure which is a good result for a big set of labels, and relatively small learning data.

Output information obtained from the CRF system is similar to the annotation given in Figure 6.20. So, it indicates a fragment of text (the first and subsequent tokens) that contains information relevant to an assigned label. Thus, we know which attribute is assigned to a text fragment but we don't know its value. For an attribute with not many values, like diabetic type (D_TYPE) which has three values, it is possible to learn these values, too. We have to add a value to a label and to learn a full label consisting of an attribute and a value, e.g. (D_TYPE|TYPE_1). If there are many possible values, like for example for COMP (complication) attribute that has 17 different values, or for numerical values, it

Table 6.4: Results of the labels assignment for the model 4

Concept	F-score	Precision	Recall	Ref. annot.	CRF annot.
recommendation_str	100	100	100	160	160
ID_P_SEX	100	100	100	96	96
creatinin_str	100	100	100	182	182
W_IN_WORDS	100	100	100	15	15
ACET_D	100	100	100	85	85
ACC_DISEASE	100	100	100	48	48
ID_AGE	100	99	100	223	225
I_TYPE	99	99	100	797	804
HEIGHT	99	99	100	78	79
dose_str	99	100	99	1,364	1,356
id_str	99	100	98	183	180
DIET_TYPE	99	98	100	86	88
hospit_str	99	100	97	173	168
WEIGHT	99	98	99	135	137
H_FROM	98	98	97	304	301
insulin_treat_str	98	96	100	404	420
H_TO	98	98	98	302	302
D_TYPE	98	97	99	141	144
COMP	97	97	97	294	293
meals_str	97	98	96	94	92
cal_str	96	98	95	130	125
diet_str	96	96	96	79	79
D_TREAT	96	92	100	45	49
BMI	96	93	98	113	119
N_COMP	96	100	92	96	88
HBA1C	95	93	98	254	268
feature_l_str	94	93	96	188	193
EDUCATION	94	97	91	229	216
D_CONTROLL	93	90	96	159	170
ORAL_TREAT	93	93	93	97	97
DIET_CORRECTION	92	100	86	7	6
insulin_inf_treat	92	86	100	6	7
THERAPY_MODIFF	91	84	100	62	74
FROM_IN_W	91	100	83	12	10
cure_l_str	90	86	93	214	232
DOSE_MODIFF	88	85	92	25	27
RELATIVE_DATA	88	83	94	52	59
lipid_str	75	73	77	596	624
microalbuminury_str	73	58	100	33	57
reaso_l_str	73	62	90	165	240
LDL1	63	47	96	28	58
I_THERAPY_BEG	–	0	0	1	6

is necessary to write an additional program to extract values from indicated text fragments.

The results achieved show that the carefully designed rule based information extraction system can be successfully used for preparing training data for the machine learning IE approach. Even using data that was not manually corrected, we obtained a labelling model good enough for recognising complex labels.

Terminology extraction

The history of terminology as a scientific field started in the 1930s. The foundation of traditional terminology was developed by Eugen Wüster in his doctoral dissertation. He developed the principles of terminology creation and standardisation, and collaborated in establishing Infoterm — The International Information Centre for Terminology in 1971, among many other projects.

The word ‘terminology’ is ambiguous. The Miriam-Webster dictionary includes two meanings:

1. the technical or special terms used in a business, art, science, or special subject;
2. nomenclature as a field of study.

The same difference exists in the Polish language; see, for example, “Słownik języka polskiego” Doroszewski (1969):

1. *ogół terminów, którymi posługuje się dana dziedzina wiedzy, techniki; mianownictwo*
‘all the terms that are used by a branch of science, technology; nomenclature’;
2. *nauka o terminach*
‘a scientific discipline concerning terms’.

The origin and meaning of the Polish words *termin* ‘term’ and *terminologia* ‘terminology’ are discussed in the paper by Grucza (1991a). It provides a history of terminology as a scientific field and its scientific status between the linguistics and the domain from which the terms originate. Moreover, the authors indicate the importance of terminology in contemporary scientific life as a method of communication. The Grucza paper is an introduction to other papers presented in the book “Teoretyczne podstawy terminologii” (ang. Theoretical Foundations of Terminology) (Grucza, 1991b) where, among others things, the structures of terms in English, French and German are described by Wojnicki (1991) and the problem of normalisation of Polish terminology is discussed by Rybicka-Nowacka (1991).

The problem of terminology as a scientific discipline, its theory and history is described in the book written by Maria Teresa Cabré (1999). The author highlights the interdisciplinary character of terminology from the socio-linguistic point of view and describes its relationship to these fields and to the logic. She considers terminology as an independent scientific discipline (p. 32):

“...terminology is an interdisciplinary field of enquiry whose prime object of study are the specialized words occurring in natural language which belong to specific domains of usage.”

Juan C. Sager, in the book “A Practical Course in Terminology Processing”, outlines the theoretical foundation of terminology, its linguistic dimension and its role in the field of communication. The author describes the connection of terminology to computer science and covers compilation, storage and retrieval of terminology. He denies the existence of terminology as a separate scientific field with the following declaration (page 1):

“This book denies the independent status of terminology as a discipline but affirms its value as a subject in almost every contemporary teaching programme... We see terminology as a number of practices that have evolved around the creation of terms, their collection and explication and finally their presentation in various printed and electronic media. Practices, however well established, do not constitute a discipline..”

Regardless of doubts as to whether terminology is a separate scientific discipline or not, it is connected to many scientific disciplines and is an area of research for computational linguistics. The importance of the latter field is noted by Maria T. Cabré in the following sentence (page 6):

“Computer science is one of the most important forces behind changes in terminology.”

In this chapter, we describe the way computer science can help in the development of terminology resources. First, we present a method for obtaining terminology from a domain corpora. We discuss problems that need to be resolved in this task. Then we present several approaches to the problem of Automatic Term Recognition (ATR).

7.1 Domain terminology as a set of terms

The concept of domain terminology is widely understood and used. Many domain terminology dictionaries have been created for all languages, for Polish, e.g. economic (Gęsicki and Gęsicki, 1996), literary (Głowiński *et al.*, 2010), and fine art (Kozakiewicz, 2014). These dictionaries contain terms and their definitions, and sometimes examples of usage. Let us now consider the notion of a term. Sager (1990) gives the following definition (page 19):

“The items which are characterised by special reference within a disciplines are the ‘terms’ of that discipline, and collectively they form its ‘terminology’”

On pages 89–90, the author states the principles which terms should fulfil, but he points out that the following criteria are highly idealistic:

1. The term must relate directly to the concept. It must express the concept clearly. A logical construction is advisable.
2. The term must be lexically systematic. It must follow an existing lexical pattern and if words are of foreign origin, a uniform transcription must be preserved.
3. The term must conform to the general rules of word-formation of the language which will also dictate the word order in compounds and phrases.
4. Terms should be capable of providing derivatives.
5. Terms should not be pleonastic (i.e. no redundant repetition, e.g. combining a foreign word with a native word having the same meaning).
6. Without sacrificing precision, term should be concise, and should not contain unnecessary information.
7. There should be no synonyms, whether absolute, relative or apparent.
8. Terms should not have morphological variants.
9. Terms should not have homonyms.
10. Terms should be monosemic.
11. The content of terms should be precise and not overlap in meaning with other terms.
12. The meaning of the term should be independent of context.

In the computer science area, a domain term, i.e. an element of domain terminology, is usually defined as a word or a phrase which represents a concept from the domain. Unfortunately this description does not include any hints on how a domain term might be recognised in text. Our rough definition of a term is as follows:

“a term is a nominal phrase which is used in the domain texts frequently enough to make it plausible that it represents something important and it does not occur equally frequently in texts on different subjects”.

This definition is based on the frequency notion which might be easy to implement in a computer program, but is difficult for a human being to apply. Experts preparing domain terminology tend to prepare it not only on the basis of texts but use their domain knowledge. Therefore, they probably classify the term ‘myocardial infarction’ as a medical domain term regardless of how many times it occurs in data. It may appear only once in data concerning treatment of children and quite often in everyday news as a cause of death of famous people. So, if we focus on the frequency of phrases in texts to automatically obtain a list of terms, the quality of the result, i.e. the set of terms, is strongly correlated to the quality of data — we usually obtain the set of terms that characterise collected texts.

The general method of terminology extraction is common for various domains and languages. Differences certainly arising from grammar features of particular languages, which affect, among other things, preparation of domain texts, construction of term candidate phrases, and recognition of different term variants. There are quite a lot of studies concerning this topic, not only for well-researched languages like English, French and German, but also for other lan-

guages, e.g. Norwegian (Øvsthus *et al.*, 2005), Bulgarian (Koeva, 2007), Lithuanian (Grigonytė *et al.*, 2011), and Chinese (Ji *et al.*, 2007), (Yang *et al.*, 2008). For well-researched languages, several commercial and publicly available tools exist. Links to some of them are available in the *Terminology extraction* article on Wikipedia and in the Terminology Coordination Service created in 2008 (<http://termcoord.eu/>) by the European Union to stimulate and coordinate terminology works. For Polish, only a few papers have been published concerning collocation recognition: (Buczyński and Okniński, 2005), (Broda *et al.*, 2008), (Pęzik, 2012), (Pęzik, 2014), and papers concerning economic terminology identification (Marciniak and Mykowiecka, 2013) and medical terminology recognition (Marciniak and Mykowiecka, 2014b).

Automatically extracted terminologies are useful in many scientific fields. They are valuable for the creation of glossaries, vocabulary or terminological dictionaries, the creation of resources for document annotation, automatic translation, ontology and construction of knowledge databases. Each task requires a slightly different approach to the creation of resources. Usually, the differences lie in various assumptions concerning construction of extracted phrases. A general method for the terminology extraction task and common problems can be found in the next section.

7.2 Terminology extraction process

An automatic terminology extraction procedure usually consists of the four steps shown in Figure 7.1. It starts from collecting domain texts — a source of domain terminology. Texts are pre-processed with basic linguistic tools to segment them into tokens, and perform morphological analysis and disambiguation. As a result, we obtain data segmented into tokens and word forms annotated with their base form, part of speech, and morphological feature values.

The next step identifies candidates for terms. It has to be decided what types of phrases are considered as terms. Terminology usually consists of noun phrases, but, in some approaches, verbal phrases are also taken into account (Savova *et al.*, 2003). In most approaches, a list of term candidates is established on the basis of linguistic information and rules of phrase construction. The English phrases might be schematically defined by the following regular expressions that define linguistic filters for term construction:

$$((A|N)^* \text{Prep?}) (A|N)^* N$$

or its simplifications like:

$$(A|N)^* N$$

Where A and N stand for adjective and noun, respectively.

To extract term candidates, shallow grammars are most often defined (Frantzi *et al.*, 2000), but there are also solutions where this task is carried out on fully syntactically parsed texts, see (Salton, 1988), (Savova *et al.*, 2003). The candidates for domain terms can also be simple n-grams, e.g. (Wermter and Hahn,

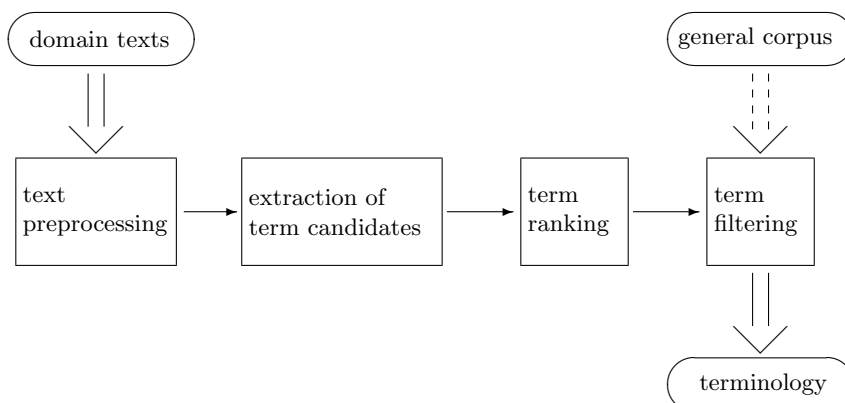


Fig. 7.1: Data processing stages

2005), (Pantel and Lin, 2001), out of which those fulfilling the grammatical requirement for terminology phrases are selected.

To create term candidates, a ‘stop-list’ is frequently used to prevent an extraction of candidates which probably shouldn’t be taken into account as terms. These are, among others, words relating to a context, like ‘next’, or ‘highest’. Terms including them are in contradiction with the last criterion of Sager (1990) given in Section 7.1, as concepts modified with such adjectives are identifiable only in the context, e.g. ‘the next bus stop’ or ‘the highest building’. A stop-list should include complex prepositions and their parts, e.g. ‘in case of’, ‘with reference to’; in Polish, their equivalents are *w przypadku*, and *w odniesieniu*. If such expressions are not excluded, we create the following terms: *przypadek podwyższonego ciśnienia* ‘case of high-pressure’ from a part of recommendation *w przypadku podwyższonego ciśnienia...* ‘in the case of high-pressure...’.

The next step consists of term ranking which is the essential issue in any terminology extraction task. The aim of this step is to find the best way of ordering term candidates into a list that consists of one-word and multi-word terms relevant to the set of collected domain documents. Methods combine different term features like term frequency, term length, co-occurrence strength of words, and diversity of contexts in which terms occur, see Section 7.4. Kageura and Umino (1996) introduce two notions into discussions about automatic term recognition:

unithood: the degree of strength or stability of the syntagmatic linguistic unit.¹

A high unithood is typical for compound words, collocations, idiomatic expressions and terms;

termhood: the degree to which a lexical unit is related to a domain-specific concept, or, in other words, it represents a domain concept.

¹ Hisamitsu and Tsujii (2003) describe it as: “togetherness of words”.

The unithood approach is based on the methods developed for collocation recognition. Pecina and Schlesinger (2006) summarise many lexical association measures, based on e.g.: mutual information, statistical tests of independence, and likelihood measure. Some researchers, see Korkontzelos *et al.* (2008) or Paziienza *et al.* (2005), claim that in the biomedical domain, termhood-based methods outperform unithood-based methods. But these methods are related and, to some degree, difficult to differentiate. For example, there is no consensus on the classification of the most widely used ATR technique in recent years, i.e. the C-value method described in Section 7.4.1, which is based on frequency measure, and number of term contexts. Nakagawa and Mori (2003) classified the C-value method as “basically based on unithood”, while most researchers treat it as a termhood method see (Paziienza *et al.*, 2005), Milios *et al.* (2003). Zhang *et al.* (2008) describe the same method as “hybrid approach, in which ‘unithood’ and ‘termhood’ are combined to produce a unified indicator”, Fedorenko *et al.* (2013) classify the method as hybrid, too. We think that these doubts are due to the difficulty with classification of measures based on ordinary frequency into one of the discussed types, as they might be considered as unithood measures as well as termhood ones.

Applying a ranking procedure we get an ordered list of the potential terms. We expect that phrases which are not domain relevant or linguistically incorrect are located low on this list. A cut-off value according to a coefficient value or a position on the list is chosen at the next processing stage. A set of phrases which are located above this cut-off constitute a terminology list.

The final step of automatic terminology extraction is filtering out non-domain terms. It is usually done on the basis of comparison of phrases obtained from domain and non-domain corpora. Damerau (1993) propose using ratios of relative frequency between a domain corpus and a general one, to identify phrases characteristic for the domain, see also (Manning and Schütze, 1999). Chung and Nation (2004) propose, in the paper concerning terminology extraction from a technical corpus, classifying one-word terms as domain (technical) terminology when they appear 50 times more often in the domain corpus than in a general one. Tests showed that the comparison corpus would need to be at least around 2,000,000 words long. These results are not supported by larger studies, so their usefulness to similar experiments is not clear. Gelbukh *et al.* (2010) propose a method based on comparison with a general corpus using log-likelihood similarity for recognition of one-word domain terms.²

Several approaches are inspired by a method used for Information Retrieval, namely the TF-IDF measure (Term Frequency–Inverse Document Frequency, see theoretical justifications by Robertson (2004)) that indicates how important a term is to a document within a document collection. Due to the importance of the method in many applications, we will describe it in more detail. The measure is defined in Equations 7.1–7.3 for which t is a term, d is a document, D is a set

² In this solution, Gelbukh *et al.* select one-word domain terms as candidates for creating multi-word terms. So the step of filtering out non-domain terms is at the beginning of the ATR procedure.

of documents, and N is the total number of documents in the set D . $TF(t,d)$ is the frequency of the term t in the document d , and is defined in Equation 7.1.

$$(7.1) \quad TF(t, d) = \frac{n_{t,d}}{\sum_{i \in d} n_{i,d}}$$

where $n_{i,d}$ is the number of occurrences of the term i in the document d .

Inverse document frequency says that if a term is common or rare across documents, it is counted as logarithm of the total number of documents in the set D in which the term t appears divided by the number of documents in D . It is defined in Equation 7.2, where 1 is added to the denominator to avoid division by a zero value if the term t is not present in the set of documents D .

$$(7.2) \quad IDF(t, D) = \log \frac{N}{1 + |\{d \in D : t \in d\}|}$$

TF-IDF is calculated as in Equation 7.3.

$$(7.3) \quad TF-IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

High TF-IDF(t,d,D) indicates that the term t is specific for the document d in the considered collection of documents D , i.e. it occurs many times within a small number of documents. If we substitute the set of documents D with the set of domain corpora, we can apply the method to indicate terminology that is specific to a considered domain. The method and its modifications are used in automatic term recognition, not only to filter out non-domain terms but also as a useful measure for term ranking algorithms, see Section 7.4.3.

7.3 Variants recognition

An important constraint concerning the terminology formulated by Sager (see Section 7.1) is that terms should describe unique concepts. In other words, terminology should not contain synonyms. This requirement imposes the need for standardisation of terms; so different variants realising the same concept should be recognised and substituted by one form — the canonical form.

This problem is thoroughly discussed for the biomedical domain in (Nenadić *et al.*, 2002), (Nenadić *et al.*, 2004), (Ananiadou *et al.*, 2006). Terminology normalisation for this domain is more complex than for other domains such as economics or humanities due to tokenisation problems (see Section 5.5), plenty of acronyms and abbreviations, and orthographic variants. So, the solutions proposed for this domain might equally well be applied to another one. In this section, we briefly present their approach to the problem. They itemise the following types of normalisation:

Orthographic — it concerns spelling variants such as:

- differences between British and American English spelling;

- differences in spelling words originating from Latin and Greek (often in the medical domain);
- inconsistent usage of hyphens and slashes in terms, e.g. ‘T-cell’ and ‘T cell’.

Morphological — it covers:

- normalisation of inflectional forms; in English it comes down to recognition of phrases in the singular and plural, and forms in the genitive (‘Down’s syndrome’ and ‘Down syndrome’), but in languages with rich inflection this problem is more complex, see Section 8.2;
- phrases with derivative words, e.g. ‘cellular gene’ and ‘cell gene’ in English, in Polish: *jama brzuszna_{adj}* and (rarely) *jama brzucha_{noun}* both mean ‘abdominal cavity’.

Lexical — it concerns synonyms, i.e. words or phrases that have the same or nearly the same meaning in one domain, like ‘flu’, ‘influenza’ and ‘grippe’ in English or *guz* and *tumor* ‘tumour’ in Polish.

Structural — it concerns phrases that have the same meaning but have different syntactic structure such as:

- in Polish adjectives, they may appear before and after a noun and don’t usually change their meaning, e.g. *lewa nerka*, *nerka lewa* ‘left kidney’;
- in English phrases with or without preposition, e.g. ‘viral infection’, ‘infection of viral’;
- term coordination, e.g. ‘adrenal glands and gonads’ or *usg tarczycy i jamy brzusznej* ‘ultrasound of thyroid and abdominal cavity’ that should be divided into two simpler terms *usg tarczycy* ‘ultrasound of thyroid’ and *usg jamy brzusznej* ‘ultrasound of abdominal cavity’.

Acronyms and abbreviations of the same word or phrase should refer to their expanded forms. For example, in Polish medical documents all the following acronyms *RM*, *MRI*, *NMR*, *MR* have their expanded form *rezonans magnetyczny* ‘magnetic resonance’.

Ananiadou *et al.* (2006) use semantically isomorphic transformations to reduce phrases to their canonical forms. For orthographic variants, they use a set of transcriptions, where term candidates are mapped to their normalised forms. The mapping is prepared manually. For example, ‘oe’ is transformed into ‘e’ so ‘oestrogen’ has its normalised form ‘estrogen’.

For acronym recognition, they scan data for retrieval of acronym definitions. In order to do this, they prepared patterns that extract several types of acronym definitions. An acronym may precede or follow its expanded form, see examples (Nenadić *et al.*, 2002):

- ‘(e.g. tumour necrosis factor alpha TMF-alpha)’;
- ‘(MIBP Myc-intron-binding peptide)’.

Park and Byrd (2001) take into account text markers and a list of cue words like *stands/short/acronym/for short* that indicate the existence of an abbreviation and an abbreviation definition. The next step after definition extraction consists in matching expanded forms with acronyms. Usually, it is a combination of the

initial letters. Different ways of acronym creation are summarised by Chang *et al.* (2002) in Table 7.1. Ananiadou *et al.* (2006) propose heuristics for recognising omissions or insertions of some letters in phrases to create acronyms.

Table 7.1: Ways of acronym creation, (Chang *et al.*, 2002)

Acr.	Definition	Description
VDR	vitamin D receptor	The letters align to the beginnings of the word.
PTU	propylthiouracil	The letters align to a subset of syllable boundaries.
JNK	c-Jun N-terminal kinase	The letters align to punctuation boundaries.
IFN	interferon	The letters align to some other place.
SULT	sulfotransferase	The abbrev. contains contiguous characters from a word.
ATL	adult T-cell leukemia	The long form contains words not in the abbreviation.
CREB-1	CRE binding protein	The abbrev. contains letters not in the long form.
beta-EP	beta-endorphin	The abbreviation contains complete words.

As it is possible to use several acronyms for one concept, it is necessary to link those representing one concept. The problem of acronym recognition for English is discussed in many papers, see (Park and Byrd, 2001), (Chang *et al.*, 2002), (Okazaki and Ananiadou, 2006). The main difference among methods of acronym recognition and normalisation is in the way the best is chosen, i.e. the most suitable extension from the set of collected phrases (definitions). It is usually done with help of statistics, but in some approaches (Park and Byrd, 2001), this task is performed by rules with priorities. They are established on the basis of information as to how often a rule is applied to find an extension of an abbreviation. Priorities are updated during the operation of the system.

We are not aware of any automatic system for abbreviation recognition for Polish. Note, however, that in the biomedical domain, Polish abbreviations are quite often created from English phrases. For acronyms: *RM*, *MRI*, *NMR*, *MR*, only the first one is the acronym of the Polish phrase *Rezonans Magnetyczny* ‘magnetic resonance’ while the rest of the acronyms are created from English phrases: MR — Magnetic Resonance, MRI — Magnetic Resonance Imaging, NMR — Nuclear Magnetic Resonance. So, this task should take into account texts in both languages as sources of abbreviations, at least for domains where foreign acronyms are popular.

Lexical variants are usually recognised with the help of synonym dictionaries. In order to recognise the structural variation of terms, a set of transformation rules was prepared by Nenadic *et al.* (2004). For example, to change a term with

the ‘of’ preposition into a phrase without the ‘of’ preposition they apply the following rule:

if structure of term candidate **is** $(A|N)_1^* N_1 \text{ Prep}(\text{of}) (A|N)_2^* N_2$
then CR= $(A|N)_2^* N_2 (A|N)_1^* N_1$

Table 7.2 quotes an example of normalisation. According to the classification above, the second phrase is a structural variant of the phrase ‘human cancer’, the third one — a morphological variant, and the last — a lexical one.

Table 7.2: A term and the corresponding canonical representation, (Ananiadou *et al.*, 2006)

Term	Canonical representative
human cancer	} human cancer
cancer in humans	
human’s cancer	
human carcinoma	

7.4 Examples of terminology extraction methods

In this section, we present five methods of terminology selection. As there are many studies on this problem, we decided to put forward only a subjective choice from a great number of them. Our criterion for preparing of this brief overview is a diversity of approaches. First, we describe the standard method in term ranking, namely the C/NC value method, which we use in our study of terminology extraction in Polish (Marciniak and Mykowiecka, 2013, 2014b). The description of the method given in the next section comes from our paper (Marciniak and Mykowiecka, 2014b).

7.4.1 C/NC method

One of the most popular ranking methods designed for recognising multi-word terms is the C/NC method proposed in Frantzi *et al.* (2000). This method takes into account phrase occurrences, both in isolation and nested inside longer ones, and the different contexts of their appearances. In this method, every phrase is assigned a C-value, which is computed on the basis of the number of times it occurs within the text, its length, and the number of different contexts it takes (within noun phrases in which it occurs). The definition of the C-value coefficient is given in Equation 7.4

$$(7.4) \quad C\text{-value}(p) = \begin{cases} l(p) * (freq(p) - \frac{1}{r(LP)} \sum_{lp \in LP} freq(lp)), & \text{if } r(LP) > 0, \\ l(p) * freq(p), & \text{if } r(LP) = 0 \end{cases}$$

p is a multi-word phrase under consideration,

LP is a set of phrases containing p ,

$r(LP)$ is the number of different phrases in LP ,

$l(p) = \log_2(\text{length}(p))$.

Long phrases tend to occur more rarely than shorter ones, so multiplication by the logarithm of length moves them towards the leading positions. If a nested phrase occurs in one context only, its C-value is set to 0 as it is assumed to be incomplete. If a nested phrase occurs in a lot of different contexts, the chance that it may constitute a domain term increases.

A popular modification of the method was aimed at extending the ranking procedure for phrases of the length 1 which originally all get a 0 value. For this purpose, the logarithm of the length for one word phrases (used in the original solution) was replaced with a non-zero constant. In Barrón-Cedeno *et al.* (2009), where this method was applied to Spanish texts, the authors initially set this constant to 0.1, but finally set it to 1, arguing that, otherwise, one word terms would be located too low on the ranking list.

The C-value obtained using the equation cited above reflects only the relationships between the terms themselves. The results can be improved on the basis of the contexts in which the terms occur within texts. In Frantzi *et al.* (2000) it was suggested that appearing in the same context as highly ranked terms should increase the rank of the candidate term. The idea is implemented by the NC coefficient which is counted according to the following equation in which t is a candidate term, C_t is a set of distinct contexts of t , $f_t(b)$ is the frequency of b occurring as a context of t and $weight(b) = t(b)/n$ where $t(b)$ is the number of terms the context word b occurs with and n is the total number of the terms considered.

$$(7.5) \quad NC\text{-value}(t) = 0.8 * C\text{-value}(t) + 0.2 * \sum_{b \in C_t} f_t(b) * weight(b)$$

In the original solution contexts were just strings of wordforms surrounding the given phrase within the text. Barrón-Cedeno *et al.* (2009) proposed using lemmas of the surrounding words instead of their forms for processing Spanish, which has different forms of adjectives and nouns according to number and grammatical gender.

7.4.2 A statistical term extractor

Pantel and Lin (2001) propose a method that starts with extraction of two-word term candidates which are then extended to multi-word units with the help of the mutual information and log-likelihood measures. The method looks rather

complex and requires experimentally fixing several thresholds. It doesn't fit the general workflow presented in Section 7.2 and the result of this method is a set of terms instead of an ordered list. Below, we sketch the method as it is proposed in Pantel and Lin's paper, where the full algorithm with details is given.

Let us first define statistical metrics which are used in the method. The mutual information is defined in Equation 7.6, where 'x' and 'y' are words or terms and $p(x,y)$ is the probability of the 'x y' in the data, and $p(x)$, $p(y)$ are probabilities of 'x' and 'y' respectively.

$$(7.6) \quad MI(x, y) = \frac{p(x, y)}{p(x)p(y)}$$

Log-likelihood is defined in Equation 7.7, where 'x' and 'y' are terms, $C(x, y)$ is the frequency of the 'x y' string and '*' denotes any token.

$$(7.7) \quad \begin{aligned} \log L(x, y) = & ll\left(\frac{k_1}{n_1}, k_1, n_1\right) + ll\left(\frac{k_2}{n_2}, k_2, n_2\right) \\ & - ll\left(\frac{k_1+k_2}{n_1+n_2}, k_1, n_1\right) - ll\left(\frac{k_1+k_2}{n_1+n_2}, k_2, n_2\right) \end{aligned}$$

where $k_1 = C(x, y)$, $n_1 = C(x, *)$, $k_2 = C(\neg x, y)$, $n_2 = C(\neg x, *)$ and $ll(p, k, n)$ is defined in Equation 7.8.

$$(7.8) \quad ll(p, k, n) = k * \log(p) + (n - k) * \log(1 - p)$$

If all 'x', 'y' occur together, the mutual information for them is high, like the log-likelihood. The authors observe that the mutual information does not work well for very infrequent terms, while the log-likelihood measure is high for frequent terms that rarely occur together. Neither of these features are desired, so Pantel and Lin combine both measures in Equation 7.9, where *minMutInfo* is a constant fixed experimentally .

$$(7.9) \quad S(x, y) = \begin{cases} \log L(x, y) & \text{if } MI(x, y) \geq \text{minMutInfo} \\ 0 & \text{otherwise} \end{cases}$$

Two-word term extraction

Pantel and Lin count frequencies for all bigrams. For all 'x y' bigrams, they look for all 4-grams consisting of adjacent tokens, i.e. 'w x y z' and check if 'x y' is the strongest connection in this 4-gram. They check conditions given in Equation 7.10, where d is a fixed constant. If any of them is fulfilled, they subtract these occurrences from the 'x y' frequency.

$$(7.10) \quad \begin{aligned} MI(x, y) &< MI(w, x) - d \\ MI(x, y) &< MI(y, z) - d \end{aligned}$$

The 'x y' bigram is selected for further processing if its final frequency is greater than an experimentally fixed threshold and $S(x,y)$ is higher than a minimal fixed score. The chosen bigrams are a nucleus of multi-word terms.

Multi-word terms extraction

The authors create all possible extensions of the bigrams from the set created according to the rules described above. They consider all phrases containing bigram ‘b’ and up to a fixed number of tokens to the left and right side of ‘b’. For further consideration, they select, from all these phrases, only those that occur several times (another threshold needs to be determined).

In the next recursive step, for each selected ‘C’ phrase (consisting of shorter terms C_1, C_2) and each word ‘w’ adjacent to this phrase they check if Equation 7.11 is fulfilled for a fixed threshold t . If yes, they add ‘w’ to the G list of the good extension of ‘C’ in decreasing order of $S(w, C)$.

$$(7.11) \quad S(w, C) < S(C_1, C_2) - t$$

The extension from the G list is added to the final list of terms if it fulfils the following three conditions:

- it is not a substring of an already accepted term phrase,
- it is not recursively extended to the longer phrase,
- it is a correctly constructed phrase.

Finally, they add those two-word candidates which are not included often enough³ in three-word and longer terms to the list of terms.

7.4.3 Contrastive measure

Roberto Basili *et al.* (2001) propose a method of term ranking which takes into account the frequency of the term head element⁴ counted in several domain collections of documents. It ranks terms in one domain, taking into account its distribution in corpora from several other domains. The approach is different for one-word terms and multi-word terms, as one-word terms are more polysemic than multi-word terms, so the authors suggest to treating them differently.

The starting point is just simple frequency, as it is indicated in many studies as a very effective method. The authors extract candidates that fulfil the grammatical criteria and indicate their head elements, as the ranking of multi-word terms depends on the distribution of the head element in resources. For each domain corpus, two separate lists are created for one-word terms and for multi-word terms. Unfortunately, they also are ranked separately, so, as a result, we obtain two lists of term candidates for each domain. The interesting aspect of the method is the possibility of taking into account cross-domain statistical measures and to be able to decrease the importance of terms appearing in many domains.

³ Unfortunately, the meaning of “not often enough” it is not defined.

⁴ The head element of a phrase is the word around which the whole phrase is built. It determines its syntactic description.

One-word terms ranking

The authors count Inverse Word Frequency (IWF) for each candidate term t according to Equation 7.12. The *IWF* does not depend on any domain, it penalises frequent candidate terms without taking into account whether they are spread out over domains, or it is strongly related to the domain and all occurrences of the term appear in one domain.

$$(7.12) \quad IWF(t) = \log \left(\frac{N}{F(t)} \right)$$

where $F(t) = \sum_{i=1}^{i=j} f_i(t)$ is the cumulative frequency throughout all domains j and $f_i(t)$ is the frequency of the term t in the domain i ; N is the sum of all frequencies of term candidates in all corpora.

The weighting function for the term t related to the interesting domain i is counted as in Equation 7.13. The authors call it contrastive weight.

$$(7.13) \quad w_i(t) = \log(f_i(t)) * IWF(t)$$

where $f_i(t)$ is the frequency of term t in domain i .

Multi-word terms ranking

The measure of a multi-word term ct in a corpus i depends on the contrastive weight (Equation 7.13) of the head element of the term ct counted for i and the multi-word term frequency. The definition is given in Equation 7.14.

$$(7.14) \quad cw_i(ct) = w_i(head(ct)) * (f_i(ct))$$

where $f_i(ct)$ is the frequency of the multi-word term ct in the domain i .

7.4.4 Method using domain-specificity and term cohesion

The method is proposed in the paper (Park *et al.*, 2002) and implemented in the GlosEx tool developed at IBM Research. The aim of the tool is to create glossary items which contain text terms important to a domain. Glossary items may consist of canonical forms and their variants. They take into account several types of variants like abbreviations, inflectional variants, different spellings with or without slash, hyphen or space. They also recognise misspelled words.

The method refers to domain specificity (*TD* defined in Equation 7.15) and term cohesion (*TC* defined in Equation 7.16). It is a hybrid approach that joins termhood and unithood measures of terms to establish their ranking coefficient.

The domain specificity of a term is calculated on the basis of the domain specificity of words included in the term, while the domain specificity of a word is correlated to its probability in both text collections: domain texts and general ones. The domain specificity of a multi-word term is the average of the domain specificity of words included in it.

$$(7.15) \quad TD(t) = \frac{\sum_{w_i \in t} \log \frac{p_d(w_i)}{p_c(w_i)}}{|t|}$$

where $|t|$ is a number of words in the term t , $p_d(w_i)$ is the probability of the word w_i in domain texts and $p_c(w_i)$ is the probability of the word w_i in general texts. The probabilities are estimated by the frequencies of words divided by the size of the corpus.

To calculate the term cohesion, the authors generalise the Dice coefficient to be able to calculate this measure to longer than two-word phrases. The measure is higher for a phrase whose words tend to occur mainly in that phrase.

$$(7.16) \quad TC(t) = \frac{|t| * \log f(t) * f(t)}{\sum_{w_i \in t} f(w_i)}$$

where, as in the previous equation, $|t|$ is a number of words in the term t , $f(t)$ is the frequency of the term t in the domain texts and $f(w_i)$ is the frequency of the word w_i in the same texts.

Since one-word terms have too high a TC measure in comparison to longer terms, as it only depends on their frequency, the authors propose reducing their TC value, taking only a fraction into account (e.g 10%).

Terms are ranked according to the term confidence measure C defined in Equation 7.17.

$$(7.17) \quad C(t) = \alpha * TD(t) + \beta * TC(t)$$

where α and β are constant values establishing the contribution of TD and TC to the ranking term coefficient C , $\alpha + \beta = 1$.

7.4.5 Term extraction from sparse data

In this section, we outline the method proposed by Ittoo and Bouma (2013). The authors claim that their method can cope with two important problems of terminology extraction: silence, i.e. detecting rare terms in sparse data, and noise, i.e. recognising domain terms despite informal or ungrammatical language. To overcome POS-tagging errors or punctuation errors, they relax linguistic patterns for candidate term recognition, claiming that the next stages of the proposed method filter out erroneous phrases.

Selecting domain phrases

To extract domain specific phrases, the authors compare phrases extracted from a domain corpus with those extracted from the general, normative corpus. For the latter one they use Wikipedia. For each term candidate they estimate its probability in the domain corpus and in Wikipedia. The termhood score of a candidate term is counted as a ratio of its probabilities in both corpora. All scores are normalised to values from 0 to 500. If a candidate term is not present

in Wikipedia, it obtains a maximal termhood score, i.e. 500. For the next stage of the method, the authors accept phrases for which the termhood score is above a chosen threshold.

Term Ranking

Ittoo and Bouma use the method of counting the unithood score for two-word phrases and extend it to complex phrases. For two-word candidates they calculate the unithood using the cube mutual information measure defined in Equation 7.18.

$$(7.18) \quad \text{unithood}(x, y) = \log \frac{\left(\frac{f(x, y)}{N}\right)^3}{\frac{f(x)}{N} * \frac{f(y)}{N}}$$

where $f(x, y)$ is the occurrence frequency of the phrase 'x y', $f(x)$ and $f(y)$ are frequencies of the 'x', 'y' tokens, and N is the size of the domain corpus.

For each candidate phrase of length $n > 2$, the authors create all sub-expressions consisting of 2 to $n - 1$ elements of the phrase. These elements appear in sub-expressions in the same order as in the original phrase. For example, for the 'w x y z' candidate phrase they create the following sub-expressions:

two-words : 'w x', 'w y', 'w z', 'x y', 'x z', 'y z';
 three-words : 'w x y', 'w x z', 'w y z', 'x y z'.

They count unithood of terms in order of the increasing length of phrases, so when they count the unithood of an n -element phrase, the unithoods of shorter phrases have already been counted. The unithood of any candidate is the sum of the unithoods of its sub-expressions divided by n — the phrase length. The unithood scores are normalised in the same way as the termhood scores within a range $\langle 0 - 500 \rangle$. One word terms have the maximal unithood score equal to 500 as their unithood is perfect, see Kageura and Umino (1996). The final domain term list consists of candidate phrases which have a unithood score above a chosen threshold.

Terminology extraction from hospital documents

In this chapter, we describe the experiments with terminology extraction from Polish domain corpora. Our approach is based on the method described in (Frantzi *et al.*, 2000), see Section 7.4.1. Two features of this method attract our attention. It allows us to rank all extracted candidate phrases of various length. As a result, it creates one list with the most appropriate terms at its top part. Moreover, the method focuses on nested terms, which may not appear by themselves in data but may constitute notions relevant and important to the considered domain.

The starting point for an ATR task is an annotated domain corpus. Chapter 5 describes how to develop such a resource for Polish. To facilitate the extraction of noun phrases (candidates for terms) from a corpus, we use an auxiliary hash file (see Section 6.4.1), which aggregates all necessary information in a more compact way than the set of XML files. For recognising the selected types of nominal phrases considered afterwards as term candidates, we define the shallow grammar described in Section 8.1.2. Then, we discuss various problems connected to the application of the C-value method to Polish data, i.e. recognition of phrases in various grammatical forms, different approaches to term ranking and the problem of filtering of general terms.

The ideas, experiments and results described in the chapter have been published in the following papers:

- The paper (Marciniak and Mykowiecka, 2013) describes the application of the C-value method to the economic corpus plWikiEcono which consists of textual content of articles that have economy related headings in Polish Wikipedia and articles that are linked to them. In the paper, we introduce the idea of using simplified base forms for counting numbers of occurrences of Polish terms which have different forms depending on numbers and cases. The data contains encyclopedic information, and there are quite a lot of general phrases that should be filtered out from the result term list. We describe an approach to this problem, by comparing the terminology extracted from the domain texts with phrases extracted from the general corpus of Polish. The paper contains a manual evaluation of the top part of the terminology list as well as a comparison of the extracted terminology with the manually created economic dictionary SEJFEK (Savary *et al.*, 2012).
- The paper (Marciniak and Mykowiecka, 2014b) describes the terminology extraction from the corpus of children’s hospital discharge records (see Section 5). In the paper, we discuss different methods of counting contexts of terms and the influence of the NC method applied to the final list of terms.

The results obtained within the experiment are manually evaluated and are compared to the Polish MeSH (Medical Subject Headings).

- In the paper (Marciniak and Mykowiecka, 2015) (the extended version of the paper (Marciniak and Mykowiecka, 2014a)), we propose a method for identifying nested terms based not only on grammatical correctness, but also on Normalised Pointwise Mutual Information (NPMI). NPMI is counted for all bigrams in a given corpus. The lowest NPMI within a phrase indicates the weakest connection within it, which suggests the best place for dividing the phrase into two parts. The proposed method is tested on two Polish corpora (used in the papers described above), and for English on the GENIA corpus (Kim *et al.*, 2003).

8.1 Term phrases

8.1.1 Phrases description

The starting point for each terminology extraction task is a decision as to what kind of phrases satisfy the conditions for being a term. Usually, terminology resources consist of noun phrases. Most approaches don't include verbal constructions like *usunięto trzeci migdal* 'adenoid was removed' in terminology resources. Researchers justify the decision that these phrases are represented in the nominalised version *usunięty trzeci migdal* 'removed adenoids'. As the problem of nominalisation of verbal constructions is not straightforward and, typically, terminology resources do not include verbal phrases, we took into account only nominal phrases, too.

An internal construction of nominal phrases might be very complex. There are several grammar aspects such as agreement, word order and coordination that should be taken into account in noun phrase analysis. They are described in linguistic hand-books such as (Szober, 1953), (Bąk, 1984), and (Polański and Nowak, 2011), among others. Formal approaches to these problems are presented e.g. in (Saloni and Świdziński, 2001), (Przepiórkowski *et al.*, 2002) and (Świdziński and Woliński, 2009). As not all nominal phrases are likely to constitute terms, it is reasonable to limit them to the syntactic structures described below:

- A single noun or an acronym, e.g. *badanie* 'examination', *nerka* 'kidney', or *USG* – acronym of 'ultrasonography'.
- A noun followed or preceded by an adjective that satisfies the agreement condition, so the adjective is in the same case, gender and number as the modified noun, e.g. *nerki*_{*n,gen:f:sg*} *lewej*_{*adj,gen:f:sg*} (kidney left)¹ 'left kidney', *wysoka*_{*adj,nom:f:sg*} *gorączka*_{*n,nom:f:sg*} 'high fever'.
- A noun modified by an adjective phrase expressed by an adjective modified by an adverb, e.g. *nieco*_{*adv*} *mniejsza*_{*adj,nom:f:sg*} *objętość*_{*n,nom:f:sg*} 'slightly smaller volume'.

¹ The word for word translation is given in parenthesis.

- A sequence of a noun and another noun in the genitive (in this construction, there are no agreement constraints), e.g. *ból_{n,nom} głowy_{n,gen}* (ache head) ‘headache’. In Polish, genitive modifiers typically follow modified nouns but they can also precede them, e.g. *rodziców_{n,gen} dom_{n,nom}* is the correct construction. This construction does not occur in domain/scientific texts as it is marked and its usage is limited to belles lettres. Thus, we decided not to take it into account in our grammar identifying term phrases.
- A combination of the last three structures, with more than one of the adjectives and the genitive modifier, as illustrated in the following examples:
 - *silny_{adj,nom:m3:sg} ból_{n,nom:m3:sg} głowy_{adj,gen:f:sg}* ‘severe headache’;
 - *nerka_{n,nom:f:sg} prawa_{adj,nom:f:sg} prawidłowa_{adj,nom:f:sg}* ‘normal right kidney’;
 - *nerki_{n,nom:f:pl} prawidłowe_{adj,nom:f:sg} wielkości_{adj,nom:f:sg}* ‘kidneys of normal size’;
 - *złamanie_{n,nom} kości_{n,gen} ręki_{n,gen}* ‘hand fracture’;
 - *wodonercze niewielkiego stopnia dolnego układu podwójnego nerki prawej* ‘mild hydronephrosis of the duplicated lower collecting system of the right kidney’.
- A noun phrase modified by a coordination of noun or adjective phrases, e.g. *nawadnianie dożylnie i doustne* ‘intravenous and oral irrigation’, *zapalenie zatok i ucha środkowego* (inflammation sinus and ear middle) ‘sinusitis and otitis media’.
- A noun phrase modified by prepositional phrases, e.g. *znieczulenie bez powikłań* ‘anesthesia without complications’.

Nouns and adjectives in the above constructions can also be respectively realised by gerunds as in *ustąpienie_{ger,nom} gorączki_{n,gen}* and past participles as in *nieco_{adv} zwiększona_{ppas,nom} echogeniczność_{ger,nom}* ‘slightly increased echogenicity’. The description of appropriate word classes is given in Appendix A.

For prepositional phrases, it is difficult to decide if they are noun phrase modifiers or fulfil different roles in a sentence, i.e. they can be valency constraints. If we accept a solution where a noun phrase can have many prepositional modifiers, the phrase *Dziewczynka przyjęta do_{prep} oddziału w_{prep} trybie ostrego dyżuru z_{prep} powodu urazu palca* ‘The girl admitted to the ward under emergency department because of a finger injury’ is a candidate for a term. A similar situation is in the following sentence — *proszę zgłosić się do_{prep} tutejszej poradni w_{prep} ustalonym terminie z_{prep} aktualnymi badaniami* ‘please come to the local clinic within a fixed period with current examination results’. Both are the potential sources of semantically odd phrases like *poradnia w ustalonym terminie z aktualnymi badaniami* ‘clinic within a fixed period with current examination results’ as both prepositional phrases: *w ustalonym terminie* ‘within a fixed period’ and *z aktualnymi badaniami* ‘with current examination results’ refer to the verb *przejść* ‘come’ not to the noun *poradnia* ‘clinic’.

In the paper (Marciniak and Mykowiecka, 2013) we took into account one preposition modifier of a noun phrase. We took this decision as we wanted to compare our results with the manually prepared terminology resource which con-

tained preposition modifiers. In later works concerning medical data, we excluded from nominal phrases those containing prepositional modifiers and a nominal coordination.

8.1.2 Shallow grammar

Nominal phrases can be recognised with the help of one of the shallow parsing tools (e.g. Spejd, Gate, SProUT, Nltk tools) or one can write a dedicated program for this task, like in (Marciniak, 2010b). Each tool cooperates with its own input and output format. The general information extraction tools like GATE or SProUT accept just text as input, and perform a basic language analysis (tokenisation, morphological analysis) according to resources that cooperate with them. The output is usually an XML file. In each case, it is necessary to prepare results in a format convenient for further processing. As we have an environment for processing corpora (designed within the LUNA project), especially a tool for cascade grammar development prepared by (Mykowiecka, 2010), we decided to use it.

The tool operates on the hash file described in Section 6.4.1 that contains data annotated with part of speech and morphological features. A grammar consists of several sets of rules. The results obtained by applying one set of rules are used as the input for the subsequent set. Our shallow grammar consists of eight sets of rules being regular expressions. The grammar is used for the experiments with medical data described in (Marciniak and Mykowiecka, 2014b). The full grammar is given in Appendix C. Below, we describe constructions recognised by subsequent levels.

1. The first set of rules recognises basic elements which can be used to build up nominal phrases:
 - inflected noun elements: *subst* (noun), *ger* (gerund)';
 - noninflected noun elements: *brev:pun:np*, *brev:pun:nphr*, *brev:npun:np*, *brev:npun:nphr*, i.e. abbreviations of a noun or a noun phrase requiring or not requiring a period afterwards;
 - sequences of up to three foreign words without any additional constraints as we do not analyse the internal structure of Latin or English sequences: *foreign_subst*, *foreign* tokens;
 - inflected adjective elements: *adj* adjective, *ppas* past participle, optionally modified by a preceding adverb;
 - non-inflected adjective elements: *brev:pun:adjw*, *brev:npun:adjw*;
 - conjunction;
 - words that should be excluded from terms discussed later in this section.
2. The second set of rules recognises adjectival constructions and the most typical noun phrases:
 - consisting of adverbs after adjectives; an adverb can be attached to an adjective only when there is no adjective after the adverb as in *chłopiec leczony zachowawczo* ‘boy treated conservatively’, but in the phrase *rytm zatokowy bardzo zwolniony* (rhythm sinus very slow) ‘slow sinus rhythm’

the adverb *bardzo* ‘very’ can’t be attached to the adjective *zatokowy* ‘sinus’ as it is followed by the adjective *zwolniony*;

- construction of complex adjectives like *mózgowo-rdzeniowy* ‘cerebrospinal’ containing a special form of an adjective ending with “-o” followed by a hyphen and an adjective;
 - consisting of a noun preceded by an adjectival construction (the construction fulfils agreement constraints) or a noun followed by a non-inflected noun.
3. The third set of rules describes:
 - compound adjectival phrases consisting of a sequence of adjective phrases;
 - nominal phrases consisting of a nominal phrase followed by an adjectival phrase with fulfilling agreement constraints.
 4. The fourth set of rules combines nominal phrases with adjectival ones.
 5. The fifth set of rules:
 - recognises nominal phrases followed by nominal phrases in the genitive (or their coordination);
 - allows for recognition of a non-inflective noun as a last phrase element; it is applied, among other things, for recognition of ICD10 codes at the ends of diagnoses, e.g. *zapalenie wyrostka robaczkowego K36* ‘appendicitis K36’.
 6. The sixth set of rules combines nominal phrases with an additional nominal phrase in the genitive or an adjectival modifier.
 7. The seventh set of rules recognises coordination of nominal phrases.
 8. The eighth set of phrases indicates selected nominal phrases that make terms.

In the first set of rules we mark some words that should be ignored during the terminology phrase construction. If we include them in a terminology, it results in a subset of phrases which we consider non-domain terms. For example, from the phrase *na podstawie RTG* ‘on (the) basis (of) X-ray’ may result in an odd nominal phrase *podstawa RTG* ‘basis (of) X-ray’. These words belong to the following three classes:

- general time or duration specification, e.g. *czas* ‘time’, *miesiąc* ‘month’, *rok* ‘year’, *tydzień* ‘week’, *dzień* ‘day’, *letni* ‘years_{adj}’, *miesięczny* ‘monthly’, *roczny* ‘annual’;
- names of months, weekdays;
- introductory/intension specific words, e.g. *cecha* ‘feature’, *przyczyna* ‘reason’;
- nouns creating complex adverbs such as *na skutek* ‘through’ or a complex preposition such as *w trakcie* ‘during’.

To recognise terms that are nested inside other more complex terms, our shallow grammar adds information about an internal structure of extracted phrases. We mark the limits of substrings matched by rules applied at the subsequent levels of the grammar. We indicate only the end of a phrase and its type by the ‘>’ sign with an appropriate type. A type informs us whether a phrase may be divided in this place. The marker $>_a$ (adjective) indicates the end of an adjectival phrase. It does not allow us to divide a phrase into two nominal subphrases,

although it is still possible to shorten the phrase by cutting the indicated adjective. In the following examples we use $>_n$ for simple nouns. The markers $>_{ng}$ and $>_t$ indicate the ends of a nominal phrase recognised by different sets of rules. Both of them indicate a place where a phrase might be divided into subphrases. Examples of phrases together with their grammar output are given below:

- wynik $>_n$ badanie $>_n$
 result $>_n$ examination $>_n$
 ‘result of examination’
- karta $>_n$ informacyjna $>_a>_t$ leczenia $>_n$ szpitalnego $>_a>_t>_{ng}$
 card $>_n$ information $>_a>_t$ treatment $>_n$ hospital $>_a>_t>_{ng}$
 ‘hospital treatment information card’
- zwężenie $>_n$ ujścia $>_n$ zewnętrznego $>_a>_t$ cewki $>_n$ moczowej $>_a>_t>_{ng}$
 narrowing $>_n$ orifice $>_n$ external $>_a>_t$ duct $>_n>_t$ urethral $>_a>_t>_{ng}$
 ‘narrowing of the external urethral’
- USG $>_n$ stawu $>_n$ biodrowego $>_a>_t$
 USG $>_n$ joint $>_n$ iliac $>_a>_t$
 ‘ultrasound of the hip’

On the basis of the structural information, we can identify nominal subphrases. For example, in the second phrase we can create the following phrases (given in their nominal form) *karta informacyjna* ‘information card’, *leczenie szpitalne* ‘hospital treatment’, *karta informacyjna leczenia*, ‘treatment information card’, but markers show us that creating the phrase *informacyjna leczenia szpitalnego* is unacceptable.

8.2 Simplified base forms

All term recognition methods based on frequency of term occurrences, including the C-value method, require identification of various phrase forms. For highly inflected languages like Polish, for which different forms of a word and a phrase can vary significantly, the problem of term equality is much harder than for English. Polish nouns decline, so have various forms in different cases and numbers; adjectives additionally have different forms depending on gender and degree. For English, this problem is reduced to identification of rather regular plural noun forms and a genitive case reflected by addition of the suffix ‘-s’ or just an apostrophe. Thus, the problem of recognition different forms of the same phrase is not given much attention in related studies for English. Let us consider the phrase *pacjent z przewlekłym nieżytem żołądka* ‘patient with chronic gastritis’ in which we want to recognise occurrences of the following phrases in the nominative form:² *przewlekły nieżyt żołądka* ‘chronic gastritis’, and four nested phrases: *nieżyt żołądka* ‘gastritis’, *przewlekły nieżyt* ‘chronic inflammation’ *nieżyt* ‘inflammation’ and *żołądek* ‘stomach’. None of these five phrases can be directly identified by just matching them with the considered phrase.

² Phrases in the nominative case are traditionally considered as the basic form.

In Table 8.1, 12 forms of the phrase *przewlekły nieżyt żołądka*_{subst:gen} ‘chronic gastritis’ in both numbers and six cases³ are given. As may be noticed, almost all forms of the considered phrase are different.

Table 8.1: Declination of *przewlekły nieżyt żołądka* ‘chronic gastritis’

Case	Singular	Plural
nom	<i>przewlekły nieżyt żołądka</i>	<i>przewlekłe nieżyty żołądka</i>
gen	<i>przewlekłego nieżytu żołądka</i>	<i>przewlekłych nieżytów żołądka</i>
dat	<i>przewlekłemu nieżyтови żołądka</i>	<i>przewlekłym nieżytom żołądka</i>
acc	<i>przewlekły nieżyt żołądka</i>	<i>przewlekłe nieżyty żołądka</i>
inst	<i>przewlekłym nieżytem żołądka</i>	<i>przewlekłymi nieżytami żołądka</i>
loc	<i>przewlekłym nieżycie żołądka</i>	<i>przewlekłych nieżytach żołądka</i>

In order to unify various phrase forms in an easy manner, we operate on simplified base forms of phrases, which consist of a sequence of phrase element lemmas. We proposed this approach in (Marciniak and Mykowiecka, 2013). For the phrase in Table 8.1, the simplified base form is *przewlekły nieżyt żołądek* ‘chronic inflammation stomach’. Phrases in simplified forms allow for identification of all nested phrases (represented in simplified forms, too) by matching strings. Moreover, it is possible to join phrases that contain abbreviations with those in full forms. Of course, an abbreviation must have its expansion in an appropriate form. For example the following two phrases, with the abbreviation of the second word created ad hoc: *babka lan* and *babka lanc*, have their full form: *babka lancetowata* ‘ribwort plantain’ (it is used in patch tests). They can be unified if the base forms of the tokens *lan* and *lanc* are expanded as *lancetowaty*.

Now, let us consider the formally correct approach in which a phrase base form is in the nominative case and in the singular number (or plurale tantum, if appropriate). It means that the phrase head and adjectives modifying it have the same case, gender, and number. But, there are still several decisions that should be made to have a unique base form for a phrase. It should be decided if a nominal modifier in the genitive case has to be in the singular or plural. For example, in the following phrase: *zaplanowano operację usunięcia niewielkich torbieli* ‘removal operation of small cysts have been planned’ the nominal phrase may have two base forms: *operacja usunięcia niewielkich torbieli* ‘removal operation of small cysts’ or *operacja usunięcia niewielkiej torbieli* ‘removal operation of a small cyst’. The question is, whether these two phrases refer to the same term or are two different terms. Moreover, there are phrases whose meaning is in contradiction with the singular number of a genitive modifier like *obustronne zapalenie uszu* (bilateral inflammation ears) ‘bilateral otitis’. The phrase *obustronne zapalenie ucha* (bilateral inflammation ear) is odd. Thus, using the

³ In medical data, and many other domain corpora, the vocative case is not used, so we don’t include these forms in the table, and in considerations hereinafter.

simplified base form is easier and probably more effective than converting noun phrases to a nominative base form. The problem of creation of the nominal phrase form from the simplified one is described in Section 8.7.

Sometimes, the simplified base forms described above are treated as equivalent phrases whose meanings are different or should be considered as different (depending on assumptions concerning the term definition). This may happen due to the following reasons (Marciniak and Mykowiecka, 2014b):

- phrases with genitive modifiers occurring in different numbers, e.g. *zapalenie ucha* ‘ear inflammation’ and *zapalenie uszu* ‘ears inflammation’, have the same simplified base form: *zapalenie ucho* ‘inflammation ear’;
- the adjectives in different degrees (small, smaller) have the same base forms, e.g. *miednica mała* ‘small pelvis’ (more frequently written as *mała miednica*, where *mała* ‘small’ refers to its size) and *miednica mniejsza* (*mniejsza* ‘smaller’ indicates anatomic part) ‘lower pelvis’;
- negated and positive forms of adjectival participles, e.g. *powiększony*/*niepowiększony* ‘increased’/‘not increased’, both have the lemma *powiększyć_{inf}* ‘increase’;
- gerunds and participles have infinitives as their base forms, so phrases:
 - *usunięcie_{ger} kamienia_{subst:gen}* ‘removing stone’ — an operation,
 - *usunięty_{ppas} kamień_{subst:nom}* ‘removed stone’ — description of the stone,
 have the same simplified base form: *usunąć_{inf} kamień_{subst}*.

Hereinafter, if we compare phrases, we understand that we compare their simplified base forms. So this process unifies all grammatical forms of a phrase: *jama brzuszna* ‘abdominal cavity’ in the nominative case and *jamy brzusznej* in the genitive case are equal. For both of them, the simplified base form is *jama brzuszny*.

8.3 Candidate terms ranking

In the C-value method, the term ranking depends on a parameter related to the phrase length, its frequency in the corpus and the number of different contexts in which it appears in the corpus. The definition of the C-value method (see Section 7.4.1) leaves some room for its interpretation and, consequently, for the final ranking of terms. Let us repeat here the definition of the C-value coefficient:

$$(8.1) \quad C\text{-value}(p) = \begin{cases} l(p) * (freq(p) - \frac{1}{r(LP)} \sum_{lp \in LP} freq(lp)), & \text{if } r(LP) > 0, \\ l(p) * freq(p), & \text{if } r(LP) = 0 \end{cases}$$

- p is a phrase under consideration,
- LP is a set of phrases containing p ,
- $r(LP)$ is the number of different phrases in LP ,
- $l(p) = \log_2(\text{length}(p))$.

A parameter that may be changed in the C-value definition is how the length of a phrase affects its significance on the term list. In the C-value method, long phrases are usually promoted by multiplication of their frequency by the logarithm of the length. It moves longer phrases towards the leading positions on the ranking list. Relating the term position to the logarithm of its length means that the C-value for one word phrases has to be defined separately, if we want to include them in the terminology list. So, for one word phrases the logarithm of the length has to be replaced with a non-zero constant. Its value is responsible for a distribution of one-word terms in the ranking list. In many papers this constant is set to 0.1, and we assume this value too.

The generally accepted length coefficient for multi-word phrases is the base 2 logarithm from the phrase length. This coefficient is not supported by any scientific theory — it is fixed experimentally and we can tune it to our data. For instance, in an experiment with the terminology extraction from GENIA corpus (Marciniak and Mykowiecka, 2015) we obtained better results when the base 2 of the logarithm was exchanged by the base 4, as the average length of a term was equal to 3.8. So for GENIA corpus, it occurred that a modification assigning less weight to longer phrases gave better results.

Table 8.2: Contexts of phrase *przepuklina pachwinowa* ‘inguinal hernia’

<i>planowa</i> ‘planned’	<i>operacja</i> ‘surgery’	<i>przepukliny</i> ‘herina’	<i>pachwinowej</i> ‘inguinal’	<i>lewostronnej</i> ‘left-side’
	<i>operacja</i> ‘surgery’	<i>przepukliny</i> ‘herina’	<i>pachwinowej</i> ‘inguinal’	<i>lewostronnej</i> ‘left-side’
<i>planowa</i> ‘planned’	<i>operacja</i> ‘surgery’	<i>przepukliny</i> ‘herina’	<i>pachwinowej</i> ‘inguinal’	
	<i>operacja</i> ‘surgery’	<i>przepukliny</i> ‘herina’	<i>pachwinowej</i> ‘inguinal’	
		<i>przepuklina</i> ‘herina’	<i>pachwinowa</i> ‘inguinal’	<i>lewostronna</i> ‘left-side’
	<i>lewostronna</i> ‘left-side’	<i>przepuklina</i> ‘herina’	<i>pachwinowa</i> ‘inguinal’	
		<i>przepuklina</i> ‘herina’	<i>pachwinowa</i> ‘inguinal’	<i>prawostronna</i> ‘right-side’
		<i>przepuklina</i> ‘herina’	<i>pachwinowa</i> ‘inguinal’	<i>obustronna</i> ‘both-sides’
	<i>prawostronna</i> ‘right-side’	<i>przepuklina</i> ‘herina’	<i>pachwinowa</i> ‘inguinal’	
	<i>uwięźnięta</i> ‘incarcerated’	<i>przepuklina</i> ‘herina’	<i>pachwinowa</i> ‘inguinal’	<i>prawostronna</i> ‘right-side’

The definition of the C-value method relates to a set of contexts of a considered phrase, where by a context we mean another phrase being a term candidate and containing the phrase. The C-value of a phrase depends on the phrase frequency and the number of its different contexts and their frequency. From the total frequency of the phrase we subtract the frequency of longer phrases containing the phrase divided by the number of such different phrases. This correlation indicates that if a nested phrase has more contexts it is more important and is more likely a term. If all contexts are different, we subtract only one from the total number of occurrences. If a phrase occurs always as a nested one and only in one context, all the nested occurrences are subtracted from the total number of the phrase occurrences and the C-value is equal to a 0 value.

The number of contexts might be calculated in several ways. Let us consider an example where the term *przepuklina pachwinowa* (hernia inguinal) ‘inguinal hernia’ occurs in nine different phrases given in Table 8.2. The meaning of the longest phrase: *planowa operacja przepukliny pachwinowej lewostronnej (prawostronnej)* is ‘planned inguinal hernia surgery of the left(right)-side’.

The most straightforward definition of a context⁴ of a term contained in a phrase is: the pair of strings being the phrase parts to the left and to the right side of the term. For this definition the following phrases: *operacja przepukliny pachwinowej* ‘inguinal hernia surgery’ and *planowa operacja przepukliny pachwinowej* ‘planned inguinal hernia surgery’ constitute two different contexts for the term *przepuklina pachwinowa* ‘inguinal hernia’. In our example, we obtain ten different contexts for the term *przepuklina pachwinowa* ‘inguinal hernia’ as all phrases form different contexts according to the above definition.

However, when we analyse the first two examples in Table 8.2, a question arises as to whether they do create different contexts, as the one word context for the considered term is the same. The first phrase is longer but the word *planowana* ‘planned’ does not directly concern the term *przepuklina pachwinowa* ‘inguinal hernia’. So, another way of defining the context is: one word to the right and left of the term.

Polish adjectives may appear on both sides of a modified noun; thus in practice, *przepuklina pachwinowa lewostronna* and *lewostronna przepuklina pachwinowa* ‘left-side inguinal hernia’ create the same context. This observation inspires us to consider contexts not as pairs of words but just as a set of words.

We performed several experiments concerning that problem (Marciniak and Mykowiecka, 2014b), as we hoped that this allowed us to eliminate truncated phrases from the top part of the term list. The following three methods of context counting, described above, were tested:

1. counting pairs of left and right full context combined together;
2. counting different words on both left and right side grouped together;
3. taking into account the maximum from different left and right words’ contexts counted separately.

⁴ To compare contexts easily we use their simplified base forms consisting of word for word lemmas, see Section 8.2.

Below, we give contexts of the term *przepuklina pachwinowa* ‘inguinal hernia’ counted for the phrases in Table 8.2. To avoid translating we give them in English.

1. ‘surgery’-‘left-side’, ‘surgery’-[empty], [empty]-‘left-side’, ‘left-side’-[empty], [empty]-‘right-side’, [empty]-‘both-sides’, ‘right-side’-[empty], ‘incarcerated’-‘right-side’;
2. ‘surgery’, ‘left-side’, ‘right-side’, ‘both-sides’, ‘incarcerated’;
3. ‘surgery’, ‘left-side’, ‘right-side’, ‘incarcerated’ (there are one more left than right contexts).

The best results, for the task described in (Marciniak and Mykowiecka, 2014b), we obtained for the third option that was the most restricted from the above methods of counting contexts, i.e. it gave the smallest number of contexts.

8.4 Results

The C-value method we applied to the data from two wards of a children’s hospital: the allergies and endocrine ward (further referred to as *o1*) and the surgical ward. They consisted of about 78,000 tokens, and over 360,000 tokens, respectively. Parts of this section together with the presented results were published in (Marciniak and Mykowiecka, 2014b).

8.4.1 Statistics

The grammar given in Appendix C together with the procedure for nested phrase recognition identified 4,156 different nominal phrases (nested or independent) in the *o1* set, 11,354 in the surgery set and more than 14,156 in both sets combined together. 1,354 phrases occurred in both sets (about one third of the smaller set).

The number of phrases extracted using the shallow grammar and the distribution of their length and frequencies are given in Table 8.3 and 8.4. The last column in Table 8.3 indicates how many phrases are common to both data sets. About 20% of these phrases are singular words; the largest group of phrases has two elements (38%) while only about 5% have 5 or more words. The average phrase length is equal to 2.5. More than half of the phrases occurred exactly once, while less than 10% of them occurred more than 10 times.

Table 8.5 shows the distribution of the C-value ranking according to left and right contexts grouped together (see Section 8.3). About one third of phrases got a 0 value because they occurred only as nested phrases in one context.

An advantage of the C-value method is the ability to identify terms as phrases that never occur in isolation. This feature is especially important for relatively small domain corpora, where a chance that a term doesn’t occur in isolation is high. One of the examples of the successful recognition of such a term in our medical data is *kość ramienna* ‘humerus’. Another example is *miedniczka nerki* ‘renal pelvis’ which also did not occur in isolation but had 15 occurrences in 6

Table 8.3: Distribution of phrase lengths

Phrase length	Data set			Common	
	o1	surgery	o1+surgery	nb	% from o1 in surg.
\sum	4,156	11,354	14,156	1,354	32.58
1	1,381	2,219	2,880	720	52.14
2	1,644	4,212	5,403	453	27.55
3	801	2,941	3,605	137	17.10
4	242	1,301	1,511	32	13.22
5	68	476	534	10	14.71
>5	20	205	223	2	10.00
Max	12(8)	5(7)	12(8)	0	-

Table 8.4: Distribution of phrase frequencies

Phrase freq	Data set		
	o1	surgery	o1+surgery
\sum	4,156	11,354	14,156
=1	2,272	7,120	8,211
2-10	1,417	4,076	4,572
11-50	325	922	969
51-100	71	115	157
101-1000	71	168	217
1000-	0	28	30

Table 8.5: C-value distribution

Terms freq	Data set		
	o1	surgery	o1+surgery
\sum	4,156	11,354	14,156
C=0	1,110	3,458	4,163
C>0	3,046	7,896	9,993
0<C<1	893	1,509	1,936
C=1	565	1,301	1,708
C>1	1,588	5,086	6,349
1<C<=2.5	898	2,842	3,531
C>2.5	690	2,244	2,818

different contexts and was located in 705th place. However, the strategy of promoting nested phrases on the basis of the occurrences of the phrases they are part of, can sometimes lead to undesirable results. The phrase *infekcja dróg* ‘tract infection’ never occurred alone but had 11 different contexts and was located very high (216) in spite of being an incorrect (truncated) phrase. An extreme example

of such a phrase which gained a very high C-value is *karta informacyjna leczenia* ‘treatment information card’ being a subsequence of the phrase *karta informacyjna leczenia szpitalnego* ‘hospital treatment information card’. In surgical data it occurred 1,164 times in this phrase and once in a longer phrase *poprzednia karta informacyjna leczenia szpitalnego* ‘previous hospital treatment information card’. For the C-value counting algorithm this meant that there were two different contexts in which this phrase appeared. So, the phrase was ranked in sixth top position of the term list.

Sometimes, truncated phrases turn out to be a problem when we accept terms being nested phrases which never occur in isolation. We tested several methods for counting contexts, in order to locate such phrases at the end part of a candidate term list. The best solution, named as C_1 , relies on counting super phrases which differ only in the words adjacent to a given term, and take into account the maximum from different left and right words’ contexts counted separately. It is the third method, the most restricted one, described in the previous section.

Table 8.6: C_1 -value distribution

Terms freq	Data		
	<i>o1</i>	surgery	<i>o1</i> +surgery
\sum	4,156	11,354	14,156
$C_1=0$	2,843	4,140	4,933
$C_1 > 0$	2,843	7,214	9,223
$0 < C_1 < 1$	775	1,243	1,625
$C_1=1$	581	1,339	1,757
$1 < C_1 \leq 2.5$	843	1,487	3,227
$C_1 > 2.5$	644	2,068	2,614

The distribution of the C_1 -value is given in Table 8.6. For the C_1 -value method the phrase: *karta informacyjna leczenia* ‘treatment information card’, which occurred only as the nested phrase and has only one context, obtained a proper 0 C_1 -value. The proposed strategy, however, did not eliminate all “unfinished” phrases and yielded only a slight lowering of their scores, e.g. from 28th place down to 45th for *USG jamy* ‘USG of cavity’ in the list for surgical data. The high ranking of this phrase on the terminology list is a result of it being a part of the following two phrases: *USG_{brev:nw} jamy_{subst:gen} brzusznej_{adj:gen}* (used 377 times alone and 51 as a nested phrase) and less common *USG_{brev:nw} jamy_{subst:gen} brzucha_{subst:gen}* (used 3 times alone). Both phrases have the same English equivalent: ‘USG of abdominal cavity’. Moreover, the phrase *USG jamy* was recognised once in isolation because of a spelling error in the word *brzusznej* ‘abdominal’.

C_1 coefficients are by definition usually lower than the original C-values. However, the changes in the ranking order are not very large. For *o1* data, 20

out of 600 top elements received a C_1 -value equal to 0. Only two of them were good medical terms, the rest were incomplete phrases like the one described above and were correctly suppressed. For surgical data, these extreme changes were even smaller — 4 out of 600 top phrases got a 0 C_1 -value, one of them is a correct medical term. In the entire surgical data, 119 terms which had a non-zero C-value got a 0 C_1 -value and 46 of them were incorrect phrases. For the previously given example, *infekcja dróg*, we got 4 contexts instead of 11 and the coefficient value was lowered by about 20%, but the position changed only by 20. Similarly, for the very frequent phrase *USG jamy* the change, equal to about 40% of coefficient value, resulted in a small change of 17 positions.

8.4.2 C/NC phrase reordering

As we wanted to test the impact of the NC reordering of terms, we ranked candidate terms according to the C_1 /NC method. Tables 8.7–8.8 show the leading terms for both data sets. To check if the changes introduced by the NC correction method were significant, we used the top 300 as the set of terms whose contexts were taken into consideration while calculating the NC coefficient. Unfortunately, clinical notes mostly contain noun phrases and a lot of terms just have punctuation marks as their contexts. Thus, reordering phrases according to the NC values did not introduce many changes. In fact, most corrections only caused a difference of no more than 20 places. The bigger differences were seen only at the bottom of the list where they are not very important. The end of the list is not taken into account as a source of domain terms. The possible explanation of this minor positive effect is the relatively small size of the data.

8.4.3 Manual annotation

This test was aimed at checking the completeness of the initial list of all considered nominal phrases. It involved the manual identification of terminology in documents and checking how many of these terms were present in the full list of terms before truncating it. The *o1* documents were approximately two times longer, so we randomly selected two *o1* documents (1,667 tokens) and four surgery documents (2,074 tokens) for the evaluation.

The manual evaluation was performed by two annotators: one was a paediatrician specialising in allergology and pulmonology, the second was involved in the experiment, had a computer background and had experience in linguistic and medical data processing. The two annotators were only given very general instructions to mark a phrase which they thought of as being important in clinical data and which did not include prepositions.

The results are given in Tables 8.9–8.10. As is evident from the information in the tables, about 85% of the phrases indicated by the annotators are common for both of them. The lists of extracted terms contain more than 80% of the phrases indicated by the annotators.

Analysing the results of this evaluation task, we noticed that sometimes only the boundaries of the phrase indicated by the annotators were different, e.g., in

Table 8.7: Top 20 phrases in *o1* data

Phrase	C ₁ /NC	Full	Nested
<i>karta informacyjna leczenia szpitalnego</i> 'hospital treatment information card'	185.60	116	0
<i>morfologia krwi</i> 'full blood count'	124.00	155	4
<i>wynik badania</i> 'examination result'	114.04	118	27
<i>masa ciała</i> 'body mass'	107.82	122	17
<i>stan ogólny</i> 'general condition'	102.66	75	62
<i>układ kielichowo-miedniczkowy poszerzony</i> 'widened pyelocalyceal system'	102.17	55	0
<i>pediatria ogólna</i> 'general paediatrics'	93.60	117	0
<i>oddział alergologii</i> 'allergy ward'	93.60	117	0
<i>kod pacjenta</i> 'patient code'	92.80	116	0
<i>USG jamy brzusznej</i> 'ultrasound of the abdominal cavity'	92.14	66	10
<i>lekarz prowadzący</i> 'attending physician'	91.28	114	0
<i>ordynator oddziału</i> 'head of hospital department'	91.28	114	0
<i>badanie ogólne</i> 'general examination'	79.51	93	9
<i>RTG klatki piersiowej</i> 'chest X-ray'	78.14	52	12
<i>nerka prawidłowej wielkości</i> 'kidney of normal size'	74.81	59	0
<i>pęcherzyk żółciowy prawidłowy</i> 'normal gall bladder'	73.54	58	0
<i>układ kielichowo-miedniczkowy</i> 'pyelocalyceal system'	69.35	4	59
<i>pęcherz moczowy wypełniony</i> 'filled bladder'	62.56	42	11
<i>klatka piersiowy</i> 'chest'	58.80	1	87
<i>badanie</i> 'examination'	55.20	35	665

the phrase *na całym ciele* 'on the whole body' only *ciało* 'body' was recognised by the first annotator, while the second annotator included the word *całe* 'whole'. Moreover, both annotators had a tendency to indicate phrases that contained coordinations of nouns which were not covered by the grammar, e.g. *Wyniki podstawowych badań morfotycznych i biochemicznych krwi i moczu* 'The results of basic morphotic and biochemical blood and urine examinations'. The first annotator recognised 42 terms in the *o1* data that were absent from the automatically prepared list for the following reasons: lack of grammar rules recognising the coordination of nominal phrases – 6 errors; lack of other grammar rules – 8; tagging errors – 11; problems with rules containing abbreviations and their tagging – 10; phrases containing time expressions and introductory/intension specific words (e.g. 'week', 'goal', 'direction') – 6.

Table 8.8: Top 20 phrases in surgical data

Phrase	C ₁ /NC	Full	Nested
<i>karta informacyjna leczenia szpitalnego</i> 'hospital treatment information card'	1,862.40	1,164	1
<i>oddział chirurgiczno-urazowy</i> 'surgical and casualty ward'	1,332.80	833	0
<i>badanie ogólne</i> 'general examination'	1,030.95	1,170	112
<i>wynik badania</i> 'examination result'	964.56	1,167	43
<i>oddział chirurgii</i> 'surgical ward'	943.26	1,179	3
<i>kod pacjent</i> 'patient code'	931.20	1,164	0
<i>zalecenie lekarskie</i> 'medical recommendation'	924.80	1,156	0
<i>zastosowane leczenie</i> 'applied treatment'	735.22	919	1
<i>odpływ pęcherzowo-moczowodowy</i> 'vesicoureteral ureter'	678.09	124	317
<i>pęcherz moczowy</i> 'bladder'	662.48	325	525
<i>wskaźnik protrombinowy</i> 'prothrombin ratio'	609.60	762	1
<i>stan ogólny dobry</i> 'good general condition'	526.40	414	0
<i>grupa krwi</i> 'blood group'	520.80	649	4
<i>USG jamy brzusznej</i> 'ultrasound of the abdominal cavity'	511.34	377	51
<i>układ kielichowo-miedniczkowy</i> 'pyelocalyceal system'	508.30	67	267
<i>karta informacyjna</i> 'information card'	470.00	1	1,173
<i>wsteczny odpływ pęcherzowo-moczowodowy</i> 'vesicoureteral reflux'	468.70	238	14
<i>leczenie szpitalne</i> 'hospital treatment'	466.40	0	1,166
<i>stan ogólny</i> 'general condition'	430.81	222	422
<i>nerka prawidłowej wielkości</i> 'kidney of normal size'	410.84	324	1

Table 8.9: Table 8 - Phrases in *o1* texts

	1st annot.	2nd annot.	Common
Nb of phrases	241	235	208
Nb of extr. phr.	199	190	175
% of extr. phr.	82.5	80.0	84.1

8.4.4 Manual evaluation

The second test indicated how many medical phrases were at the top, in the middle and at the bottom of the lists of terms ordered from the highest to the lowest score of their C₁/NC-value.

Table 8.10: Phrases in surgery texts

	1st annot.	2nd annot.	Common
Nb of phrases	163	164	138
Nb of extr. phr.	134	136	116
% of extr. phr.	82.2	82.9	84.0

Due to disparities in the length of candidate term lists, we evaluate terms in a ratio 1:2 reflecting this difference. For the evaluation experiment for the *o1* data, we took the top 200 terms and randomly selected 100 terms from the middle of the list ($C_1/NC\text{-value} \in (1.0, 2.5)$) and 100 from the bottom part of the list ($C_1/NC\text{-value} \in (0.0, 1.0)$). For surgery data, we evaluated the 400 topmost terms and 200 terms from the middle and bottom part of the lists. Then, the phrases were judged by the same two annotators, as to whether they belonged to the terminology or not. Not all phrases from the top part of the lists were classified as terms. Despite attempts to eliminate semantically odd phrases like *USG jamy* ‘USG of cavity’ and *infekcja dróg* ‘infection of tract’ (only in the *o1* data), they still appear in the top part of the lists as they are frequent in the data, and ‘cavity’ and ‘tract’ are part of several well established phrases. Another problem was caused by abbreviations attached to correct phrases like *uraz głowy S* ‘head injury S’ where S is a part of the ICD-10 code of the illness ‘S00’ written with a space between ‘S’ and ‘00’. Our grammar does not exclude such constructions as it is possible that an abbreviation is at the end of a phrase, e.g. *kontrolne badanie USG* ‘control ultrasound examination’.

The results of the evaluation are given in Table 8.11–8.12. In the top part of the lists, the great majority of terms (about 88%) are judged to be domain related by both annotators. The percentage of badly structured terms is below 10%. The proportion of badly structured terms in the other two sets is evidently higher, which proves that the C/NC ranking method moves bad terms toward the end of the list. However, as can be seen, even the last section of the list contains 60–82% of domain terms.

Table 8.11: Phrases considered as terms in *o1* documents

	1st annotator						2nd annotator					
	Domain		General		Bad		Domain		General		Bad	
	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%
Top200	176	88	19	9.5	5	2.5	178	89	14	7	8	4
Middle100	88	88	5	5.0	7	7.0	83	83	8	8	9	9
End100	75	75	18	18.0	7	7.0	82	82	10	10	8	8

Table 8.12: Phrases considered as terms in surgery documents

	1st annotator						2nd annotator					
	Domain		General		Bad		Domain		General		Bad	
	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%
Top400	353	88.3	28	7.0	19	4.7	348	87.0	27	6.7	25	6.3
Middle200	136	68.0	11	5.5	43	21.5	145	72.5	14	7.0	41	20.5
End200	127	63.5	33	16.5	40	20.0	121	60.5	35	17.5	44	22.0

8.4.5 Comparison with Polish MeSH

One of the methods that might be used to roughly test results obtained by an ATR method is their comparison with a domain terminology resource. In the medical domain there are many appropriate resources for English, and several other languages like French and German, designed and maintained by the U.S. National Library of Medicine within the Unified Medical Language System (UMLS <http://www.nlm.nih.gov/research/umls/>). For Polish, the only resources developed consistent with UMLS are Medical Subject Headings (MeSH), available for research purposes after registration. The data is being created in the GBL (Central Medical Library) in Warsaw and contains 26,852 main headings and 18,392 synonyms (in data from the year 2013). MeSH is a controlled biomedical vocabulary that was created to index articles from biomedical journals and to make literature searches easier.⁵

The MeSH data contains terms used for the purpose of indexing journal articles and books, such as ‘kidney’ or ‘gallbladder’, but does not contain the phrases ‘left kidney’ or ‘normal gallbladder’, as journal papers are not indexed by ‘left kidney’ or ‘normal gallbladder’. On the other hand, the latter terms are frequent in hospital documentations. In the past, experiments in applying MeSH to clinical data were done for English (Cooper and Miller, 1998) and Swedish (Kokkinakis and Thurin, 2008), while UMLS resources were used for information extraction in French (Pereira *et al.*, 2008; Hoste *et al.*, 2011), German (Markó *et al.*, 2003), and Dutch (Hoste *et al.*, 2010). A source of data that contains clinical terminology is SNOMED but it is not translated into Polish.

As there are no other publicly available electronic resources of Polish medical terminology, we compared the results obtained in the task with the terminology represented in the Polish MeSH thesaurus. The comparison of the extracted terminology in the simplified base forms with the thesaurus that contains terminology in the nominative base forms is not straightforward. There are three possible solutions.

The first one is to convert the terminology from simplified base forms into correct grammatical phrases and check them in MeSH. We have to take into account that the general Polish morphological dictionary does not recognise

⁵ For Polish, another MeSH-like resource exists that we tried to take into account in our research, but we failed to obtain access to it.

about 18.8% of word-tokens in clinical data, see (Marciniak and Mykowiecka, 2011b). The problem of the automatic generation of correct base forms from the simplified ones in general is prone to errors but the construction of medical phrases is more restricted than in literary language, so the results are better. We performed this task with the help of phrases extracted from real data, in which we identified fragments that are stable like genitive complements. This solution significantly reduces the role of unknown words. For example, in the phrase *wirus Epsteiną_{gen}-Baar_{gen}* ‘Epstein–Barr virus’, the part *Epsteiną_{gen}-Baar_{gen}* has the same form in all inflected forms of the whole phrase. So, it is possible to copy this part from the phrase extracted from real data. We have to take into account that part of the terminology in Polish MeSH is nominal phrases in plural, e.g. the above phrase is in plural form in MeSH: *Wirusy Epsteiną_{gen}-Baar_{gen}* ‘Epstein–Barr viruses’. This problem can be overcome by generating both singular and plural forms.

The second method for comparing our terms to the MeSH thesaurus consists in converting MeSH data into simplified base forms. This method also has disadvantages, as 42% of words contained in MeSH are not represented in the general Polish dictionary that we used for the annotation of our data and was used to annotate the NKJP corpus. Converting MeSH terminology into simplified base forms does not solve all problems either. For example, Polish MeSH does not contain the phrase: *chirurgia naczyniowa* ‘vascular surgery’ but it contains *zabiegi chirurgiczne naczyniowe* ‘vascular surgery operations’. The English equivalent of the last phrase contains the first phrase, but this is not true of the Polish version. The simplified form of the first phrase *chirurgia_{subst} naczyniowy_{adj}* is not contained in the simplified version of the last phrase *zabieg chirurgiczny_{adj} naczyniowy_{adj}* as the strings *chirurgia* and *chirurgiczny* are different.

The third approach is to compare the simplified forms with data in MeSH using approximate string matching. To apply this method, we perform a sort of stemming by removing suffixes indicating cases of nouns and adjectives. Then we apply the Levenshtein distance measure which takes into account the position of a non-matching letter in the analysed word. Words are more similar if differences are found nearer to the end of the word than to the beginning. For each word from a phrase in question, we find a set of similar words. Then, we look for MeSH terms that contain one similar word for each phrase element.

We tested all three methods of the third approach to perform a comparison of the top ranked surgical ward terminology with the MeSH thesaurus. As we wanted to test only medical terminology, we selected 353 terms that underwent positive manual verification by the first annotator. First, we converted terms into their correct base forms. 11 terms out of 353 contained words which were unknown to the morphological dictionary and which were supposed to be inflected: *urodynamiczny* ‘urodynamic’, *przypęcherzowy* ‘paravesical’, *de-tromycynowy* ‘chloramphenicol’ and *podpęcherzowy* ‘bladder outlet’ and compound words *pęcherzowo-moczowy* ‘vesicoureteral’ (4 terms) and *miedniczkowo-moczowodowy* ‘pelvi-ureteric’ (3 terms). The base forms of these terms were corrected manually. 52 terms (15%) are present in the MeSH thesaurus in their

exact form, while 90 (25.5%) exact forms are nested in other terms or are equal to MeSH terms. Approximate string matching performed on the simplified forms increased the number of recognised terms to 106 (30%). 9 terms recognised by the method using exact forms were not recognised by the last method. Almost all these phrases contain gerunds whose lemma forms differ significantly from the words, e.g. *leczenie szpitalne* ‘hospital treatment’ has the simplified base form *leczyć szpitalny*. Finally, we tested the approximate string matching method on the corrected base form set of terms. In this case, 119 (34%) terms give positive results.

Our results are slightly worse than the results discussed in the paper (Masarie and Miller, 1987). In that experiment from 1987, manually extracted terminology from hospital documents was compared with the English MeSH. The authors concluded that about 40% of these phrases were present in MeSH.

8.5 NPMI driven recognition of nested terms

As we wanted to eliminate incorrect phrases from the top part of the terminology list, we analysed results obtained in the terminology extraction task performed on two hospital wards (described in Section 8.4). It showed the following reasons for problems:

- Improper morphosyntactic analysis of data caused by:
 - lack of medical resources like dictionaries including medical vocabulary, acronyms and abbreviations;
 - tagging errors, as Polish taggers were trained on general language texts;
 - lack of some constructions in the shallow grammar (e.g. better coverage of coordinating structures).
- Semantically odd, truncated phrases, like *USG jamy* ‘USG of cavity’, were included in the terminology list with high scores.
- Non-domain terminology, especially one-word terms, were included in terminology resources.

On the basis of the error analysis, we proposed a method preventing the creation and promotion of truncated nested phrases to be considered as terms. The idea presented in this chapter and some parts of text were published in the paper (Marciniak and Mykowiecka, 2015).

The main idea is to use a unithood-based method, e.g. Normalised Pointwise Mutual Information (NPMI, Bouma, 2009) for driving recognition of nested phrases. Our solution is based on the division of each considered phrase into a maximum two parts. The parts into which a phrase is divided must create nested phrases consistent with grammar rules, or must create one nested phrase and its adjectival modifier. Usually, there are several possible places for division of a phrase, from which we choose the weakest point according to NPMI counted for bigrams on the basis of the whole corpus. So, as a bigram constitutes a strong collocation, it prevents the phrase from being divided in this place, and usually stops the creation of semantically odd nested phrases.

In this experiment, we took into account the corpus consisting of all data from the children’s hospital.

8.5.1 Motivations

The original C-value method recommends that all grammatical phrases created from the maximal phrases identified in a corpus should be considered as term candidates. But using this method, we obtain several nested grammatical sub-phrases which are syntactically correct, but semantically odd.

In Polish, many basic medical terms consist of two words. By basic notions we understand such concepts that may not be divided into simpler elements, e.g. *pęcherzyk żółciowy* ‘gallbladder’, *wyrostek robaczkowy* ‘appendicitis’, *staw kolanowy* ‘knee-joint’, *blona śluzowa* ‘mucous membrane’. Many of them are translated into one-word English terms. If we cut off an essential piece of such phrase we may get a grammatically correct but semantically odd phrase like *zapalenie pęcherzyka* ‘inflammation of bladder’ from *zapalenie pęcherzyka żółciowego* ‘inflammation of gall bladder’ or simpler ‘cholecystitis’. If we cut off the adjective from *staw kolanowy* ‘knee-joint’, we obtain the more general concept *staw* ‘joint’, but if we do the same for the phrase *operacja lewego stawu kolanowego* (surgery left joint knee) ‘surgery of the left knee-joint’, we obtain an odd phrase *operacja lewego stawu* ‘surgery of the left joint’. Its strangeness comes from the fact that it contains less important information that the surgery is related to the left side of the body but does not include more important information about the type of the joint.

Let us consider another example. From the phrase *infekcja górnych dróg oddechowych* (infection upper tract respiratory) ‘infection (of the) upper respiratory tract’, it is possible to create an odd phrase *infekcja górnych dróg* ‘infection (of the) upper tract’. The original phrase has many different longer phrases in which it is nested, e.g. *(częsta, drobna, ostra, bakteryjna...) infekcja górnych dróg oddechowych* ‘(often, minor, acute, bacterial...) infection (of the) upper respiratory tract’, but it always concerns *drogi oddechowe* ‘respiratory tract’. We observe that the bigram *drogi oddechowe* ‘respiratory tract’ constitutes a strong collocation. So, the original phrase shouldn’t be divided in this place to create a phrase containing the word *drogi* ‘tract’ without adding its type, i.e. *oddechowe* ‘respiratory’ in this case.

Nominal phrases are usually constructed from two parts, except for coordinated phrases and nouns with more complex subcategorisation frames, which usually do not fulfil agreement constraints in Polish. So, for a terminology nominal phrase, we suggest a division into two parts exactly and the best place for the division is indicated by the weakest bigram.

8.5.2 Algorithm

From several methods for counting the strength of bigrams we chose the normalised pointwise mutual information proposed by Bouma (2009), as it is less

sensitive to occurrence frequency. We were looking for a method in which the bigram, consisting of a rare and a frequent token, will be high if the rare token mainly appears together with the frequent token, as, for example, for *esowate skrzywienie* ‘S-shaped curvature’. The definition of this measure for the ‘x y’ bigram, where ‘x’ and ‘y’ are lemmas of sequence tokens, is given in Equation 8.2, where $p(x,y)$ is a probability of the ‘x y’ bigram in the considered corpus, and $p(x)$, $p(y)$ are probabilities of ‘x’ and ‘y’ unigrams respectively.

$$(8.2) \quad NPMI(x, y) = \left(\ln \frac{p(x, y)}{p(x)p(y)} \right) / - \ln p(x, y)$$

A candidate term list consists of all maximal phrases extracted from a corpus and nested terms identified within maximal phrases by our new method. It is based on two aspects: grammatical correctness and normalised pointwise mutual information (NPMI) counted for all bigrams on the basis of the corpus. We use NPMI to recognise the weakest points within phrases to suggest the best place for division of a phrase into only two parts.

Technically, the process is as follows: first, we extract all the grammatical phrases from the corpus, taking into account only the maximal one. We count NPMI for all bigrams in the corpus. Then, for each phrase, we call the recursive procedure `candidate_term`, whose pseudocode is given in Figure 8.1. The procedure starts from adding its argument, i.e. the phrase, to the candidate term list, if it is a valid noun phrase. Then, if it is a multi-word phrase, we identify all places where the phrase can be divided according to the grammar rules, and for these places we indicate the weakest connection according to the NPMI value. The phrase is divided into two parts at the weakest connection point, and for both fragments of the phrase, we perform the `candidate_term` procedure.

candidate_term (phr)

if phr is a valid noun phrase add phr to the list of terms

if length(phr) > 1

find all i positions where phr can be divided according to the grammar

for all i positions

count NPMI(i-th bigram of phr)

sort NPMIs from the lowest to the highest value

j := position with the lowest NPMI

divide phr into phr1 and phr2 on j-th position

candidate_term(phr1)

candidate_term(phr2)

Fig. 8.1: Initial procedure for nested phrases recognition

After considering examples of nominal phrases in Polish, we realised that the weakest connections are usually between two nominal phrases. So, adjectives are more likely interpreted as they modify the nearest noun and not the whole

phrase, as in: *prawidłowa_{adj} mikroflora_{noun} górnych_{adj} dróg_{noun} oddechowych_{adj}* ‘normal microflora (of the) upper respiratory tract’. In this phrase, this solution is correct. All the outermost adjectives are important parts of nominal phrases constructed around their nearest nouns, and it should be divided into two nominal phrases: *prawidłowa mikroflora* ‘normal microflora’ and *górne drogi oddechowe* ‘upper respiratory tract’. However, it is not the universal rule. Let us consider another example: *częste infekcje górnych dróg oddechowych* ‘frequent infections (of the) upper respiratory tract’, where *częste* ‘frequent’ modifies the whole phrase. To take into account the above specificity of adjectives in Polish nominal phrases, we introduce a slight modification to the basic algorithm, see Figure 8.2. If the weakest connection prefers the cutting of an adjective part from a phrase, we find the nearest place where the phrase is divided into two nominal phrases. Then, we compare the NPMI value referring to this bigram with 120% (fixed experimentally) of the lowest NPMI value. If it is still lower, we cut off one outermost element (adjective or adverb) from this adjectival part of the phrase, and otherwise, we divide the original phrase in that second place into two nominal phrases.

candidate_term (phr)

```

if phr is a valid noun phrase add phr to the list of terms
if length(phr) > 1
  find all i positions where phr can be divided according to the grammar
  for all i positions
    count NPMI(i-th bigram of phr)
  sort NPMIs from the lowest to the highest value
  j := position with the lowest NPMI
  if the j-th position divides phr into two nominal phrases
    divide phr into phr1 and phr2 on j-th position
  else
    n := position with the lowest NPMI where phr is divided into
      two nominal phrases
    if (120% NPMI(j)) > NPMI (n)
      divide phr into phr1 and phr2 on n-th position
    else
      divided phr into phr1 and phr2 on j-th position
  candidate_term(phr1)
  candidate_term(phr2)

```

Fig. 8.2: Procedure for nested phrases recognition

8.5.3 Examples

A good example illustrating our method is its application to the phrase *infekcja górnych dróg oddechowych* ‘infection (of the) upper respiratory tract’. Let

us compare nested phrases obtained from it with the help of the two following methods: creating all grammatically correct nested phrases, and the NPMI driven method. The considered phrase is constructed according to the following pattern:

Noun_j Adj_i Noun_i Adj_i

where indexes indicate agreement constraints, so a grammatically correct phrase may consist of: Noun_j Adj_i Noun_i, but can't be constructed as: Noun_j Adj_i. Thus, *infekcja górnych dróg* 'infection of the upper tract' is grammatically correct, while *infekcja górnych* 'infection of upper' is not. The phrase can be divided in one of two places indicated by the '|' character:

<i>infekcja</i>		<i>górnych</i>	<i>dróg</i>		<i>oddechowych</i>
‘infection’		‘upper’	‘tract’		‘respiratory’

Six grammatically correct phrases can be created from the phrase, see Table 8.13. Applying our method, we first count NPMI for the places of possible divisions. The NPMI values for two bigrams *infekcja górny* 'infection upper' and *droga oddechowy* 'tract respiratory' counted for the medical corpus are given in Table 8.14. The lower value is for the first bigram, so the phrase can be divided into: *infekcja* 'infection' and *górne drogi oddechowe* 'upper respiratory tract'. Both parts constitute nominal phrases so the phrase is divided in this place and both parts are added to the list of term candidates. In the next step, only the second phrase can be recursively divided into: *górne drogi* | *oddechowe* 'upper tract' 'respiratory' and *górne* | *drogi oddechowe* 'upper' 'respiratory tract'. The weaker connection is for: *górny droga* 'upper tract'. So, the adjective *górne* 'upper' is cut off the phrase and only the nested phrase *drogi oddechowe* 'respiratory tract' is accepted as a term candidate. Table 8.13 contains a comparison of the nested phrases obtained by both methods for the considered phrase. It may be noted that our method, correctly, does not extract two semantically odd nested phrases from the six obtained by the first method: *infekcja górnych dróg* 'infection (of the) upper tract' and *górne drogi* 'upper tract'.

Let us consider a phrase where the lowest NPMI indicates division into an adjective and a nominal phrase: *boczne_{adj} skrzywienie_{noun} kręgosłupa_{noun}* 'lateral curvature (of the) spine'. The phrase can be divided in both places: *boczne* | *skrzywienie* | *kręgosłupa* 'lateral | curvature | spine'. The weakest connection is for the bigram: *boczny skrzywienie* 'lateral curvature'. It indicates division into the nominal phrase *skrzywienie kręgosłupa* 'curvature (of the) spine', and the adjective *boczne* 'lateral'. The other place of division causes the phrase to be divided into two nominal phrases. So, we compare the NPMI for *skrzywienie kręgosłup* 'curvature spine', with 120% NPMI *boczny skrzywienie* 'lateral curvature', see Table 8.15. As the first value is lower than the second one, the method prefers to divide the phrase into two nominal phrases *boczne skrzywienie* 'lateral curvature' and *kręgosłup* 'spine'. The basic algorithm (without multiplying NPMI values, in some cases by 120%) creates a good term *skrzywienie kręgosłupa* 'curvature (of the) spine' instead of two nominal phrases: *boczne skrzywienie* 'lateral curvature' and *kręgosłup* spine.

Table 8.13: The results of two methods of nested phrases recognition for ‘infection (of the) upper respiratory tract’

The grammatically correct subphrases				NPMI driven subphrases			
‘infection’	‘upper’	‘tract’	‘respiratory’	‘infection’	‘upper’	‘tract’	‘respiratory’
<i>‘infekcja’</i>	<i>‘górnny’</i>	<i>‘droga’</i>	<i>‘oddechowy’</i>	<i>‘infekcja’</i>	<i>‘górnny’</i>	<i>‘droga’</i>	<i>‘oddechowy’</i>
<i>infekcja</i>	<i>górnnych</i>	<i>dróg</i>	<i>oddechowych</i>	<i>infekcja</i>	<i>górnnych</i>	<i>dróg</i>	<i>oddechowych</i>
<i>infekcja</i>	<i>górnnych</i>	<i>dróg</i>		—			
<i>infekcja</i>				<i>infekcja</i>			
	<i>górne</i>	<i>drogi</i>	<i>oddechowe</i>		<i>górne</i>	<i>drogi</i>	<i>oddechowe</i>
	<i>górne</i>	<i>drogi</i>		—			
		<i>drogi</i>	<i>oddechowe</i>			<i>drogi</i>	<i>oddechowe</i>
		<i>drogi</i>				<i>drogi</i>	

Table 8.14: The NPMI value for the bigrams of the phrase ‘infection (of the) upper respiratory tract’

Fragment	Bigram	Translation	NPMI
<i>infekcja górnnych</i>	<i>infekcja górnny</i>	‘infection upper’	0.65658
<i>górnnych dróg</i>	<i>górnny droga</i>	‘upper tract’	0.78773
<i>dróg oddechowych</i>	<i>droga oddechowy</i>	‘tract respiratory’	0.95089

Table 8.15: The NPMI value for the bigrams of the phrase: *boczne skrzywienie kręgosłupa* ‘lateral curvature (of the) spine’

Fragment	Bigram	Translation	NPMI	120%
<i>boczne skrzywienie</i>	<i>boczny skrzywienie</i>	‘lateral curvature’	0.67619	0.81143
<i>skrzywienie kręgosłupa</i>	<i>skrzywienie kręgosłup</i>	‘curvature spine’	0.80151	

There are a few cases when the phrase division driven by the NPMI value prefers cutting off an adjective in the first step instead of dividing it into two nominal phrases, e.g. *okoloporodowe_{adj} uszkodzenie_{noun} splotu_{noun} ramiennego_{adj} prawego_{adj}* ‘perinatal damage (of) right brachial plexus’. Despite the fact that *okoloporodowe uszkodzenie splotu ramiennego* ‘perinatal damage (of) brachial plexus’ is a good term, we would prefer the division into two nominal phrases *okoloporodowe uszkodzenie* ‘perinatal damage’ and *splot ramienny prawy* ‘right brachial plexus’. The last division reflects the internal construction of the phrase that might be important in an ontology construction task, which is one of the intended uses of the method. Then, we want to recognise a relation that joins concepts represented by nested phrases. Please note that, despite the method preferring division into two nominal nested phrases, it still (correctly) cuts off the adjective *częsty* ‘frequent’ from the phrase *częste infekcje górnnych dróg oddechowych* ‘frequent infections (of the) upper respiratory tract’.

Let us consider phrases extracted from the medical corpus with more than one possible point of division into grammatically correct parts and those placed among the 100 top positions ranked by the C-value method without the NPMI modification. We identify eight such phrases listed below. For each phrase, we indicate the division preferred by the NPMI method and show the phrases which are created in the first division step (we don't indicate phrases recognised in subsequent recursive steps of the algorithm). Moreover, we indicate phrases which are not created from the currently analysed phrase due to the use of the NPMI driven method for nested phrase recognition.

1. The phrase with splitting points at possible positions:

karta informacyjna | leczenia | szpitalnego

'card' 'information' 'treatment' 'hospital'

'hospital treatment information card';

Division preferred by NPMI: *karta informacyjna | leczenia szpitalnego*;

Supported phrases: *karta informacyjna* 'information card', *leczenie szpitalne* 'hospital treatment';

Unsupported phrase: *karta informacyjna leczenia* 'information card of treatment'.

2. The phrase with splitting points at possible positions:

wynik | badania | dodatkowego

'result' 'examination' 'additional'

'additional examination result';

Division preferred by NPMI: *wynik badania | dodatkowego*;

Supported phrase: *wynik badania* 'examination result';

Unsupported phrase: *badanie dodatkowe* 'additional examination'.

3. The phrase with splitting points at possible positions:

USG | jamy | brzusznej

'USG' 'cavity' 'abdominal'

'USG of abdominal cavity';

Division preferred by NPMI: *USG | jamy brzusznej*;

Supported phrases: *USG* 'USG', *jama brzuszna* 'abdominal cavity';

Unsupported phrase: *USG jamy* 'USG of cavity'.

4. The phrase with splitting points at possible positions:

wsteczny | odpływ | pęcherzowo - moczowodowy

'regressive' 'ureter' 'vesicoureteral'

'vesicoureteral reflux';

Division preferred by NPMI: *wsteczny | odpływ pęcherzowo - moczowodowy*;

Supported phrase: *odpływ pęcherzowo - moczowodowy* 'vesicoureteral ureter';

Unsupported phrase: *wsteczny odpływ* 'backflow'.

5. The phrase with splitting points at possible positions:
poprawa | *stanu* | *dziecka*
 ‘improvement’ ‘condition’ ‘child’
 ‘improvement of child’s condition’;
 Division preferred by NPMI: *poprawa stanu* | *dziecka*;
 Supported phrase: *poprawa stanu* ‘condition improvement’, *dziecko* ‘child’;
 Unsupported phrase: *stan dziecka* ‘child’s condition’.
6. The phrase with splitting points at possible positions:
zalecenie | *kontynuowania* | *leczenia*
 ‘recommendation’ ‘continuation’ ‘treatment’
 ‘recommendation of continuation treatment’;
 Division preferred by NPMI: *zalecenie kontynuowania* | *leczenie*;
 Supported phrase: *zalecenie kontynuowania* ‘recommendation of continuation’, *leczenie* ‘treatment’;
 Unsupported phrase: *kontynuowanie leczenia* ‘treatment continuation’.
7. The phrase with splitting points at possible positions:
zakażenie | *układu* | *moczowego*
 ‘infection’ ‘tract’ ‘unary’
 ‘urinary tract infection’;
 Division preferred by NPMI: *zakażenie układu* | *moczowego*;
 Supported phrase: *zakażenie układu* ‘tract infection’;
 Unsupported phrase: *układ moczowy* ‘unary tract’.
8. The phrase with splitting points at possible positions:
wykładnik | *stanu* | *zapalnego*
 ‘exponent’ ‘state’ ‘inflammation’
 ‘inflammation exponent’;
 Division preferred by NPMI: *wykładnik stanu* | *zapalnego*;
 Supported phrase: *wykładnik stanu* ‘exponent state’;
 Unsupported phrase: *stan zapalny* ‘inflammation’.

The fact that we do not recognise a nested phrase within a particular longer one does not necessarily result in the absence of that nested phrase in the list of term candidates or even terms. From eight unsupported phrases listed above, five are among the 500 top positions, and *stan zapalny* ‘inflammation’ occupies the high 654th position. On the other hand, the nested phrase *USG jamy* ‘USG of cavity’ obtains a very low C-value equal to 1.00, so it is located in the gap between the 15,600th and 20,500th position. It results in the correct exclusion of the phrase *USG jamy* ‘USG of cavity’ from the term list. The phrase *karta informacyjna leczenia* ‘information card of treatment’ was eliminated from the candidate term list as all phrases in the data concerned ‘hospital treatment’, so our domain corpus does not support this phrase as a domain term.

8.5.4 Statistics of phrases

This section presents the statistics of phrases obtained by the application of the C-value ranking method to two sets of term candidates from the corpus containing all documents from a children hospital. The first set contains all possible phrases fulfilling the grammatical rules, *s-phrases*, while the second one, *s&npmi-phrases*, is obtained by our method of recognition nested terms driven by NPMI. It is worth noting that we take into account the context of a nested phrase contained in a longer phrase only when the nested phrase is supported by the method.

Both methods recognise different numbers of phrases. Table 8.16 gives a comparison of these numbers. Initially, 32,809 phrases are extracted. After applying our method of nested phrase recognition, we obtained about 80% of the phrases from the *s-phrases* set. The reduction concerns phrases irrespective of their frequencies within texts. The percentage of reduced multi-word phrases is also irrespective of their length, but the number of one-word phrases is the same, as they are always correctly constructed basic elements forming phrases. The analysis of a C-value distribution shows that we finally obtained much fewer phrases with a 0 C-value.

Table 8.16: The number of recognised phrases

Length	All	=1	=2	3-5	>5
s-phrases	32,809	4,918	13,442	13,984	465
s&npmi-phrases	26,671	4,918	10,420	10,929	404

Frequency	=1	2-10	11-50	51-100	101-1000	>1000
In isolation	13,304	6,776	1,506	300	415	81
s-phrases	18,572	10,417	2,461	523	704	132
s&npmi-phrases	15,210	8,296	2,002	420	625	118

C-value	0	0<c<1	1≤c<5	5≤c<10	10≤c<100	>100
s-phrases	8,946	2,500	16,891	1,804	2,312	357
s&npmi-phrases	3,428	2,508	16,652	1,672	2,074	337

In Table 8.17, the distribution of the length of the 2,000 top phrases is given. It shows that our method introduces more longer phrases and one-word phrases into the set of the top 2,000 terms. 80 two-word phrases are eliminated from the top part of the *s&npmi-phrase* list, as they were not strongly enough supported by the method. While for one-word phrases the NPMI method usually gives only a slightly lower C-value than the method creating all grammatically correct nested phrases, so the top part of the term list includes 45 more such terms.

Table 8.17: Length distribution of 2,000 recognised phrases

	All	=1	=2	=3	=4	=5	=6	=7	=8
s-phrases	2,000	416	1,001	424	117	35	5	1	1
— only	241	15	174	40	9	3	0	0	0
s&npmi-phrases	2,000	451	921	425	142	50	6	1	4
—— only	241	50	94	41	34	18	1	0	3
Common to both	1,759	401	827	384	108	32	5	1	1

8.5.5 Evaluation

We manually evaluated the top 2,000 terms obtained by both methods: *s-phrases* and *s&npmi-phrases*. Results are given in Table 8.18. This task is done by only one annotator and terms are classified into five classes:

- Medical terms, e.g. *ostry niezbyt żółdkowy* ‘acute gastritis’.
- General terms; some multi-word terms like *różne pory* ‘various times’, and many one-word terms too general to represent any medical concept like *dół* ‘bottom’, that is recognised as a part of the following terms: *dół pachowy* ‘armpit’ and *dół biodrowy* ‘fossa iliaca’.
- Incorrect phrases are divided into three groups according to the reason for the incorrectness, but sometimes it is difficult to classify the reason. Especially, it is difficult to classify if a problem arises from the shallow grammar or from annotation errors. So, incorrect phrases are classified into the three following groups:
 - Grammar deficiencies, like fragments of phrases that need prepositions: *dziewczynka skierowana* from the phrase *dziewczynka skierowana do chirurga* ‘girl referred to surgeon’. As our grammar does not take into account preposition phrase modifiers/constraints, so sometimes we recognise only a part of a phrase.
 - Annotation errors, lack of description in dictionaries or tagging errors. For example, *Lacidofil zalecenia* ‘Lacidofil recommendation’ is created from two separate phrases. ‘Lacidofil’ ends the description of the treatment in hospital while *zalecenia* ‘recommendation’ starts the description of recommendation (without any punctuation marks). The morphological interpretation of the word form *zalecenia_{subst:nom:pl}* ‘recommendations’ is ambiguous in the form *zalecenia_{subst:gen:sg}*. As the tagger preferred the second interpretation, the incorrect phrase with a genitive modifier *zalecenia* of ‘Lacidofil’ is created.
 - Truncated phrases, *infekcja dróg* ‘infection of tract’ and other examples given in this chapter.

In the top 2,000 positions, our method locates 91.15% of medical terms and while taking into account all nested subphrases we obtain 88.9% of medical terms. The improvement results purely from the significant reduction of truncated phrases in the data from 65 (3.25%) to 17 (0.85%). There is only one trun-

Table 8.18: Evaluation of the 2,000 top phrases

	Correct		Reason of incorrectness		
	medical	general	grammar	annotation	truncated
s-phrases	1,778	84	48	25	65
— only	174	7	8	3	49
s&npmi-phrases	1,823	85	48	27	17
—— only	219	8	8	5	1
Common to both	1,604	77	40	22	16

cated phrase introduced to the 2,000 top phrases by the NPMI driven method of recognition nested phrases which is not included in the top 2,000 phrases obtained without the NPMI modification. It is *operacja przeszczepienia*⁶ ‘surgery of transplanting’ located in 2,000th place in *s&npmi-phrases* data. In the *s-phrases* data it is located a bit further back — in 2,345th position. Unfortunately, the method doesn’t eliminate this phrase and the phrase is shifted slightly higher. The evaluation shows that our method has a strong tendency not to recognise odd phrases as they are nested once, and allows us to eliminate most incorrect unfinished phrases from the top part of the ranking list.

8.5.6 Three-word phrases

An analysis of the results shows that there is still room for improvement in the method. Let us consider again the phrase *wykladnik stanu zapalnego* ‘inflammation exponent’. Both bigrams in the phrase do not indicate a strong connection of words, see Table 8.19. But *wykladnik stanu* ‘exponent (of the) state’ implies the word *zapalny* ‘inflammatory’. While the phrase *stan zapalny* ‘inflammation’ appears in different contexts, too. Our NPMI method of nested term recognition indicated the division *wykladnik stanu | zapalnego*, which is in contradiction to our observation. Our intuition suggests rather not to divide this phrase. So we propose to analyse trigrams in a similar way as bigrams.

Table 8.19: The NPMI value for the bigrams of the phrase: *wykladnik stanu zapalnego* ‘inflammation exponent’

Fragment	Bigram	Translation	NPMI
<i>wykladnik stanu</i>	<i>wykladnik stan</i>	‘exponent state’	0.64802
<i>stanu zapalnego</i>	<i>stan zapalny</i>	‘state inflammation’	0.60755

To find strong connections of three elements, we propose two measures: $NPMI_L$ that indicates how strong ‘x’ is connected to the bigram ‘y,z’ and $NPMI_R$

⁶ It might be considered as a correct phrase — a type of surgery.

that indicates how strong the bigram ‘x, y’ is connected to ‘z’. The definition of these measures for the ‘x,y,z’ trigram, where ‘x’, ‘y’ and ‘z’ are lemmas of sequence tokens, is given in Equation 8.3 and 8.4, respectively, where $p(x,y)$ and $p(y,z)$ are probabilities of the ‘x y’ and ‘y z’ bigrams in the considered corpus, and $p(x)$, $p(y)$, $p(z)$ are probabilities of ‘x’ ‘y’ and ‘z’ unigrams, respectively, and $p(x,y,z)$ is a probability of the ‘x y z’ trigram.

$$(8.3) \quad NPMI_L(x, y, z) = \left(\ln \frac{p(x, y, z)}{p(x)p(y, z)} \right) / - \ln p(x, y, z)$$

$$(8.4) \quad NPMI_R(x, y, z) = \left(\ln \frac{p(x, y, z)}{p(x, y)p(z)} \right) / - \ln p(x, y, z)$$

Table 8.20: The NPMI values for the trigram of the phrase: *wykładnik stanu zapalnego* ‘inflammation exponent’

Phrase	NPMI _L	NPMI _R
<i>wykładnik stanu zapalnego</i> ‘exponent’ ‘state’ ‘inflammation’	0.87531	0.94730

In the considered phrase the NPMI_R is high, see Table 8.20. If we fix a threshold of a strong connection to 0.9, it is greater than this value. In this case, we propose not to cut off the right element of the phrase, even if the NPMIs of bigrams indicate that it should be divided in that place. Similarly, if an NPMI_L is greater than 0.9, we propose to block the possibility of cutting off the left element of a phrase. Therefore, if both NPMI_L and NPMI_R are greater than 0.9, we propose to leave such a three-word phrase without indicating the nested phrases inside it. Table 8.21 gives examples of such phrases from medical corpus.

8.6 Comparison with a general corpus

The evaluation of 2,000 top terms (see Table 8.18) shows that filtering out general terms is important in domain terminology extraction, as 4.4% of the top 2,000 terms in our experiment are recognised as general ones.

The definition of domain terms states that they occur mainly in domain corpora, or their frequency in such corpora is much higher than in any other text resources. Ideas for filtering out non-domain terms have been developed on the basis of a comparison of phrases obtained from domain and non-domain corpora, see Section 7.2.

In our task, the first question is: how big is the set of common phrases obtained from the medical corpus and the general corpus of the Polish language (i.e. manually corrected, balanced subcorpus of NKJP). Both corpora have a similar number of words — NKJP has 1M words and the medical corpus almost

Table 8.21: The NPMI values for strongly connected three-element phrases

Phrase	NPMI _L	NPMI _R	NPMI (x,y)	NPMI (y,z)
<i>biochemiczne badania wątrobowe</i> 'biochemical' 'test' 'liver' 'liver biochemical tests'	0.95875	0.94317	0.42405	0.43907
<i>obturacyjne zapalenie oskrzeli</i> 'obstructive' 'inflammation' 'bronchi' 'obstructive bronchitis'	0.91397	0.92504	0.72659	0.77083
<i>operatio plastica herniae</i> 'surgery' 'plastic' 'hernia' 'hernia plastic surgery'	0.92748	0.92311	0.99529	0.92748
<i>ośrodkowy układ nerwowy</i> 'central' 'system' 'nervous' 'central nervous system'	0.96349	0.97891	0.57095	0.56335

0.9M, while the total number of tokens in the medical data is almost two times greater. NKJP has been manually corrected, so all words are described by their lemma and morphological features and this description is correct in almost all cases. The medical corpus contains a lot of unrecognised words, and its tagging contains a lot of errors.

We extract terminology phrases from both corpora, applying the same shallow grammar and rank phrases with the same methods. Then, we compare the terminology identified in NKJP and the medical resources. Table 8.22 shows how many terms are recognised in both corpora and the number of terms that have a higher C-value in the NKJP data. This comparison gives only a general overview, as the characteristics (i.e. number of recognised phrases, their length and frequency) of the compared corpora is different, despite the similar number of words.

Table 8.22: Common terms

	One-word	Multi-word	Total
All terms	2,113	464	2,577
C-value greater in NKJP	1,319	193	1,512
$C_{-v_{med}} > 3.0$ & $C_{-v_{NKJP}} > C_{-v_{med}}$	96	11	107
Top 2K medical terms & $C_{-v_{NKJP}} > C_{-v_{med}}$	16	0	16

For both corpora, only 16 common phrases have more than two words. The longest common phrase has five words: *objawy infekcji górnych dróg oddechowych* 'symptoms of upper respiratory tract infection'. The following three multi-word terms have a greater C-value in NKJP:

- *niewydolność krążeniowo-oddechowa* ‘cardiorespiratory failure’, occurred only once in the medical corpus;
- *badanie płynu mózgowo-rdzeniowego* ‘examination of cerebrospinal fluid’;
- *poczucie własnej wartości* ‘self-esteem’.

All these phrases may be considered as medical terms but each occurred only once in the medical data, so they are not representative for the corpus. They are not included in the top 1/3 candidate terms, so are below a reasonable cut-off point for any term list creation.

Multi-word terms that have a higher C-value in the NKJP data make up about 1.7% of multi-word medical candidate terms. Some multi-word terms with a higher C-value in NKJP are related to the medical domain, but most of them are located very low on the term list candidates, see Table 8.23 for the top 11 multi-word phrases with a higher C-value in NKJP and a C-value greater than 3.0 in medical data.

Table 8.23: Multi-word phrases with a C-value above 3.0 in medical data and with a higher C-value in NKJP

Phrase	Translation	Medical corpus		NKJP	
		C-value	Position	C-value	Position
<i>duży stopień</i>	‘high degree’	9.25	2,817	16.00	479
<i>jedna strona</i>	‘one side’	4.00	5,266	36.00	131
<i>członek rodzina</i>	‘family member’	4.00	5,509	10.00	963
<i>intensywna terapia</i>	‘intensive care’	3.00	6,674	4.00	3,260
<i>pani doktor</i>	‘Mrs doctor’	3.00	6,750	6.50	1,674
<i>jedna noc</i>	‘one night’	3.00	6,750	5.00	1,674
<i>pierwszy etap</i>	‘first stage’	3.00	7,051	8.00	1,281
<i>lewa noga</i>	‘left leg’	3.00	7,092	5.00	2,472
<i>podjąć decyzję</i>	‘taking decision’	3.00	7,215	8.00	1,295
<i>własna prośba</i>	‘own request’	3.00	7,238	5.00	2,505
<i>dom dziecka</i>	‘orphanage’	3.00	7,252	6.00	1,885

The set of common multi-word phrases for both corpora, with a C-value in medical data greater than 3.00 and higher than in NKJP, consists of 227 phrases. Most of them are medical ones, while 16 of them are rather general phrases that should not be included in a medical terminology resource. They are given in order of appearance on the candidate term list:

- *chwila obecna* ‘present moment’,
- *organizacja pożytku publicznego* ‘public benefit organisation’,
- *tutejszy szpital* ‘local hospital’,
- *tamta pora* ‘that time’,
- *trudna sytuacja* ‘difficult situation’,
- *ostatnia minuta* ‘last minute’,

- *pomoc pilota* ‘using a remote control’,
- *duży udział* ‘large proportion’,
- *liczna grupa* ‘large group’,
- *oglądanie telewizji* ‘watching TV’,
- *różny sytuacja* ‘different situations’,
- *jeden punkt* ‘one point’,
- *obowiązujący przepis* ‘existing rule’,
- *późniejszy termin* ‘later date’,
- *odrabianie lekcji* ‘doing homework’,
- *powyższa dolegliwość* ‘above ailment’.

Some of the above phrases are used to describe the behaviour of a patient (e.g. *patrzy w jeden punkt* ‘looks at one point’) and everyday recommendations like methods for turning on and watching TV in the case of epileptic patients. There are descriptions of quantities or referential expressions containing demonstrative pronouns (this, that) or referential adjectives (last, above). The latter phrases might be eliminated from the candidate term list by adding those pronouns and adjectives to the list of words that are excluded from terms at the stage of phrase recognition by a shallow grammar.

The comparison of one-word terms is much more difficult as they are very often ambiguous. There are 96 common terms with a C-value greater than 3.0 in medical data and a higher C-value in NKJP. Most of them are not medical terms like *liczba* ‘number’, *telefon* ‘phone’ or words translated differently depending on context, e.g. *gospodarka*: ‘economy’, ‘management’ or ‘farming’ occurs mainly in the phrase *gospodarka fosforanowo-wapniowa* ‘calcium-phosphate economy’ in the medical corpus. There are several words that should be included in any medical terminology as they describe: body parts: *twarz* ‘face’, *ręka* ‘hand’, *palec* ‘finger’, *usta* ‘mouth’, *dłoń* ‘palm’, *policzek* ‘cheek’, *noga* ‘leg’, *ramię* ‘arm’, and *język* ‘tongue’, or are strongly connected to health terminology: *życie* ‘life’, *zdrowie* ‘health’, *chory* ‘ill person’. Some words are slightly less associated with medical terminology like: *mowa* ‘speech’ or words used in descriptions of examination results, e.g. *cień* ‘shadow’ or *kąt* ‘angle’ (in X-ray examination). If we analyse 495 one-word terms common to both corpora and choose only those whose medical C-value is greater than 3.0 and higher than in NKJP, we can still find some questionable terms such as *numer* ‘number’ or *charakter* ‘nature’, ‘character’, *żądanie* ‘request’.

If we compare C-values for the common top 2,000 terms, we are only able to eliminate the following 16 terms: *droga* ‘tract’, *życie* ‘life’, *tysiąc* ‘thousand’, *uwaga* ‘attention’, *matka* ‘mother’, *koniec* ‘end’, *część* ‘part’, *noc* ‘night’, *klasa* ‘class’, *rodzic* ‘parent’, *początek* ‘beginning’, *wniosek* ‘conclusion’, *wartość* ‘value’, *rodzaj* ‘type’, *Warszawa* ‘Warsaw’, *mowa* ‘speech’. Some of them are important if our goal is to develop a domain ontology. For example, daytimes or family relationships are in the range of our interest.

No multi-word term has a higher C-value in NKJP within the first 2,000 medical terms.

The requirement of 50 times more occurrences in a domain corpus given in (Chung and Nation, 2004) would eliminate such medical terms as *krew* ‘blood’, *antybiotyki* ‘antibiotic’, *żołądek* ‘stomach’, *bakteria* ‘bacteria’ and many one-word terms popular in everyday language. Only a few medical words like *mocz* ‘urine’, *USG*, or *nerka* ‘kidney’ would fulfil this strong condition.

The comparison shows that there are very few phrases frequently used in both general Polish texts and hospital documents. The common multi-word phrases are usually related to medicine. The domain relatedness of one-word phrases is quite often difficult to evaluate. So, comparison of these two corpora turns out not to be substantially helpful in filtering out non-domain phrases. Perhaps we should compare the results with terminology extracted from another domain corpus instead of the corpus of general Polish.

8.7 Converting simplified forms into terminology resources

As we have written in Section 8.2, terminology ranking is done on simplified base forms, that makes recognition of various phrase forms and their contexts significantly easier. For ease of reading, we use phrase base forms consistent with Polish tradition. However, we have to remember that all lists of terms are in simplified forms. These forms are not easy to read as all nouns are given in nominative singular form and adjectives in nominative masculine form and positive degree, for example: *białko mleka krowiego* ‘cow’s milk protein’ is represented by: *białko* ‘protein’ *mleko* ‘milk’ *krowi* ‘cow’, while the simplified form for *miednica_{subst.sg.nom.f} mniejsza_{adj.sg.nom.f.comp}* ‘lower pelvis’ is: *miednica* ‘pelvis’ *mały_{adj.sg.nom.m3.pos}* ‘small’. In order to convert a simplified form of a term into its normal base form, we use the following resources:

- Morfeusz generator, which for a given base form and its part of speech creates all forms together with their descriptions.⁷
- List of all word forms together with their base form and morphological description, utilised when a word is not available in Morfeusz resources, but the appropriate form is present in the data, which is quite often the situation for medical vocabulary.
- Phrases and nested phrases with their internal structure recognised in data by the shallow grammar.

The algorithm works on a list of phrases represented by annotated elements and markers indicating phrase internal structure, see Section 8.1.2. Each phrase element is described by the following information: a word form, its base form and morphological description. So, for the phrase *atopowym zapaleniem skóry* (‘atopic’ ‘inflammation’ ‘skin’) ‘atopic dermatitis’, we have the following input:

⁷ The new version of Morfeusz (Woliński, 2014) has an option for returning only the selected forms, but it is not available in the previous version of the program used in this task.

atopowym:atopowy:adj:sg:inst:n:pos a> zapaleniem:zapalenie:subst:sg:inst:n n>
t> skóry:skóra:subst:sg:gen:f n>

Please note that the method presented below does not take into account special cases, so is not perfect but works in most cases. To create the traditional base form of a phrase, we convert an inflected part of the phrase into the nominative case and almost always the singular. So, we should change the phrase head element together with its adjectival modifiers. The rest of the phrase can be left as it is in a form recognised in the data. Terminology phrases are limited to those constructed according to the method where the first noun of a phrase is its head element, see Section 8.2. Markers indicating phrase internal structure allow us to establish the part of the phrase that probably requires a change.

First, each phrase is converted into a list containing subsequent elements, together with a description of the forms we are looking for, or the ‘bz’ marker indicating that this element should be left in the original form. For the considered example, we have to establish forms in the nominal case for the first two elements, while the last element does not need any change, see Table 8.24.

Table 8.24: Searched forms

Form	Translation	Lemma	POS	Description of a searched form
<i>atopowym</i>	‘atopic’	<i>atopowy</i>	adj	sg:nom:n:pos
<i>zapaleniem</i>	‘inflammation’	<i>zapalenie</i>	subst	sg:nom:n
<i>skóry</i>	‘skin’	<i>skóra</i>	bz	

The results of searching forms in Morfeusz is given in Table 8.25. The word *atopowy* ‘atopic’ is unavailable in that resource, which is indicated by the negative number of forms, while for the noun *zapalenie* ‘inflammation’ we have got 14 forms.⁸

The tagsets of taggers (Pantera, TaKIPI) and Morfeusz are slightly different. For example, in the Pantera output we have one neutral gender, while in Morfeusz there are two neutral genders *n1* for personal nouns and *n2* for impersonal one. We also have to remember that, in the Morfeusz generator, if one word form refers to several morphological features, they might be coded together. Let us see the form *powiększone* of the past participle *powiększony* ‘enlarged’ which represents 14 descriptions of the form in two cases: nominative and accusative and in seven genders, e.g.

powiększone, ppas:pl:nom.acc:m2.m3.f.n1.n2.p2.p3:perf

The algorithm converting simplified forms into traditional base forms first looks for a required form in the Morfeusz resources. If it is not available, we try to find it in the corpus dictionary. If it is not available there either, we indicate this fact by the exclamation mark before the missed form. For the example

⁸ In the case of nouns, we are usually looking for a form consistent with the lemma.

Table 8.25: Morfeusz results

```

atopowy#adj#sg:nom:n:pos#atopowym#
lemma: "atopowy" hom: 0 pos: "adj" -->
Liczba form#-3# /number of forms/
zapalenie#subst#sg:nom:n#zapaleniem#
lemma: "zapalenie" hom: 0 pos: "subst" -->
Liczba form#14#
1 zapalenia, subst:pl:acc:n2
2 zapaleniom, subst:pl:dat:n2
3 zapaleń, subst:pl:gen:n2
4 zapaleniami, subst:pl:inst:n2
5 zapaleniach, subst:pl:loc:n2
6 zapalenia, subst:pl:nom:n2
7 zapalenia, subst:pl:voc:n2
8 zapalenie, subst:sg:acc:n2
9 zapaleni, subst:sg:dat:n2
10 zapalenia, subst:sg:gen:n2
11 zapaleniem, subst:sg:inst:n2
12 zapaleni, subst:sg:loc:n2
13 zapalenie, subst:sg:nom:n2 /searched form/
14 zapalenie, subst:sg:voc:n2

```

presented above, it is possible to find the form of the word *atopowy* ‘atopic’ in the nominative, neutral and singular form in the corpus dictionary, so we can create the correct lemma for the considered phrase, i.e. *atopowe zapalenie skóry*. We are not so lucky in finding the appropriate form of the word *atopowy* ‘atopic’ for the following phrase in the plural: *atopowe zmiany skórne* ‘atopic skin lesions’. In this case, the correct lemma is: *atopowa zmiana skórna* ‘atopic skin lesion’ but the corpus dictionary does not include the form of the adjective *atopowy* ‘atopic’ in the nominal case, feminine gender, and singular, so the output for this phrase is: *!atopowy zmiana skórna* to indicate that the form needs to be corrected.

For the whole ‘o1’ ward, we are not able to find forms of 48 different words in about 100 phrases (some of them are nested phrases of the longer ones). This is not a lot for the total number of 4,156 phrases recognised for this data. We hope that many of these words are available in the new Morfeusz, as it allows, among other things, for creating complex words consisting of a prefix and a Morfeusz word, as for the following adjectives: *wczesnoniemowlęcy* ‘early+infantile’, *drobnobańkowy* ‘small+bubbled’ or *okołoskrzelowy* ‘near+bronchus’.

The program takes into account several exceptions for creating phrase base forms. The noun *drogi* ‘tracts’ has the singular form *droga* ‘tract’ which is available in Morfeusz. In Polish medical phrases, this word is almost only used in plural tantum: *drogi oddechowe* ‘respiratory tract’, *drogi moczowe* ‘urinary tract’,

drogi rodne ‘genital tract’.⁹ So, if the head element of a phrase is *droga* ‘tract’, we create its base form (and the base forms of modifying adjectives) in the plural. Another exception concerns the phrase *miednica mniejsza* ‘lower pelvis’, where the adjective is always in the comparative degree. The last exception we take into account is more productive as it concerns terminology phrases containing numerals before the head element, as in: *dwa moczowody*_{pl:nom:m3}, ‘two ureters’ *obie nerki*_{pl:nom:f} ‘both kidneys’ *kilka węzłów*_{pl:gen:m3} *chłonnych*_{pl:gen:m3} ‘several lymph nodes’ and *pięć posiłków*_{pl:gen:m3} ‘five meals’. The specificity of numeral constructions is described in, e.g., Saloni and Świdziński (2001).

8.8 Towards domain ontology construction

Finally, we want to write briefly about a work in progress. An application of the ATR results which attracts our attention is the automatic (or rather semiautomatic) creation of a domain ontology on the basis of a domain corpus.

We performed several experiments in this area. First, we took up the problem of clustering the most frequent terms to detect their coherent groups. We performed experiments on the medical corpus (Mykowiecka and Marciniak, 2012a) and the economic corpus plWikiEcono (Mykowiecka and Marciniak, 2012b).

For economic data, we took the 400 top terms and performed hierarchical clustering on the basis of the following contextual features:

- the strings consisting of the base forms of 1–3 tokens;
- the strings consisting of 1–3 part of speech tags;
- the base form of the nearest: verb, noun, adjective and preposition.

Moreover, we took into account sequences of coordinated terms, and several syntactic patterns in which terms appeared like: *taki jak* ‘such as’, *czyli* ‘or, that is’, *na przykład* ‘for example’, *to jest* ‘that is’ and *zarówno...jak i* ‘both...and also’ and their equivalents. Similarity coefficients were also counted on the basis of both common heads and modifiers. We took into account relations between term heads¹⁰ represented in Polish Wordnet (Piasecki *et al.*, 2009). In the experiments, we tested different weighting schemata of the coefficients used to calculate the overall similarity measure. The evaluation of the results was done on the basis of a manually prepared partition of the selected terms. The results of automatic and manual clustering were compared using the B-cubed measure (Bagga and Baldwin, 1998) with the F-score of about 0.75, which was not high, but the problem was difficult even for well-trained annotators.

The next experiment concerned clustering descriptive adjectives extracted from medical data, see (Mykowiecka and Marciniak, 2014). An evaluation of the method was done on the basis of 101 adjectives. We obtained the F-score equal to 0.664. The result showed that the method could be used to preliminary

⁹ There are medical terms with this word in the singular too, like: *droga zakażenia* ‘transmission’.

¹⁰ Only a few complex terms were then represented in the data.

identification of adjectives describing one property in big enough and coherent data. But the problem of automatic extracting ontologies from domain corpora is still open and needs careful investigation.

Summary

The book addressed the problem of constructing domain corpora and searching them for different types of information. As an example, we used corpora containing hospital discharge documents. We described how to perform analysis of such data, as well as what kind of information can be automatically obtained from it. Although we concentrated on medical data processing, the same methods can be applied to other domain data.

First, we discussed the anonymisation of hospital documents and contemporary regulations of this problem, knowledge of which is sometimes not sufficiently widespread among Polish researchers. Then, we exhaustively described problems connected with morphosyntactic analysis of noisy medical texts which should be solved to obtain the reliable annotation.

In the next chapter, we presented, in detail, the task of rule based extraction of complex information from medical texts. Three different examples of application of the rule based information extraction system results were discussed. We successfully applied them to prepare the database for quick data search, the semantic annotation of the corpus and the training data for the CRF model of information extraction. In each case, the evaluation results confirmed the high performance of this rule based method. Although the rule based approach to IE is not novel, it is well-established in the field of natural language processing. We are convinced that the method is worth presenting and considering for use on a wider scale for processing Polish data.

The book also discussed the problem of domain terminology extraction. The last chapter is devoted to the adaptation of one of the most popular methods of automatic terminology extraction (i.e. C/NC method) to Polish data. We proposed using simplified base forms for recognition of different morphological variants of terms, and tested three methods of context counting. We also proposed a modification of term candidate list creation based on normalised pointwise mutual information. We showed that our modification reduced the number of grammatically correct but semantically odd nested phrases in the top part of a list of term candidates. We also addressed the problem of filtering out non-domain terms. In the case of hospital documents, this occurred to be irrelevant, as the list of terms contained almost solely domain terminology, but it can be applied to other, more diverse, texts.

All ideas presented in Chapter 8 were tested with the help of a series of Perl scripts, which are now re-implemented as a Java tool. We hope that the first version of the TermoPL tool will be publicly available within the Clarin-PL project (<http://clarin-pl.eu/pl/strona-glowna/>) at the end of this year.

As we have written in Chapter 7, automatically extracted terms can support domain dictionary creation (including gazetteers), developing resources of domain terminology for automatic translation or annotating domain text repositories with selected terms. The most challenging task for future research is to support the construction of domain ontologies.

A

Polish tagset

In the appendix, we briefly describe a fragment of the Polish tagset. We limit the description to parts of speech (and categories assigned to them) which are used in the book. The complete NKJP tagset (and its comparison to the IPIPAN tagset) is given in (Przepiórkowski *et al.*, 2012). A summary is also available from: <http://nkjp.pl/poliqarp/help/ense2.html>.

Table A.1: Gramatical classes

Full name	Abbreviation	Example
noun	subst	<i>nerka</i> ‘kidney’
main numeral	num	<i>dwa</i> ‘two’
adjective	adj	<i>duży</i> ‘big’
ad-adjectival adjective	adja	<i>mózgowo-</i> in <i>mózgowo-rdzeniowy</i> ‘cerebro-’ in ‘cerebrospinal’
adverb	adv	<i>znacznie</i> ‘significantly’
infinitive	inf	<i>przestrzegać</i> ‘to observe’
non-past form	fin	<i>pokrywają</i> ‘cover’
gerund	ger	<i>zwolnienie</i> ‘slowing’
active adj. participle	pact	<i>stabilizujący</i> ‘stabilising’
passive adj. participle	ppas	<i>ustabilizowany</i> ‘stabilised’
preposition	prep	<i>w</i> ‘in’
coordinating conjunction	conj	<i>i</i> ‘and’
pronoun	pron	<i>ja</i> ‘I’, <i>on</i> ‘he’
abbreviation	brev	<i>r</i> (<i>rok</i> ‘year’)
particle-adverb	qub	<i>tylko</i> ‘only’, <i>już</i> ‘quite’

The morphosyntactic characteristics (see Table A.2) of the selected grammatical classes (see Table A.1) consist of the following categories:

- noun (subst): number case gender;
- main numeral (num): number case gender;
- adjective (adj): number case gender degree;
- adverb (adv): degree;
- non-past form (fin): number person aspect;
- infinitive (infin): aspect;
- gerund (ger): number case gender aspect negation;
- active adj. participle (pact): number case gender aspect negation;

Table A.2: Grammatical categories

Category	Value	Abbreviation
number	singular	sg
	plural	pl
case	nominative	nom
	genitive	gen
	dative	dat
	accusative	acc
	instrumental	inst
	locative	loc
	vocative	voc
gender	human masculine	m1
	animate masculine	m2
	inanimate masculine	m3
	feminine	f
	neuter	n
person	first	pri
	second	sec
	third	ter
degree	positive	pos
	comparative	com
	superlative	sup
aspect	imperfective	imperf
	perfective	perf
negation	affirmative	aff
	negative	neg
fullstoppedness	with full stop	pun
	without full stop	npun

- passive adj. participle (ppas): number case gender aspect negation;
- preposition (prep): case;
- pronoun (pron): number case gender person;
- abbreviation (brev): fullstoppedness,
- ad-adjectival adjective (adja), coordinating conjunction (conj), and particle-adverb (qub) do not have additional morphosyntactic descriptions.

B

Example of text annotation

In the appendix, we compare the results of the original full grammar described in Section 6.1.4 and the simplified grammar described in Section 6.3.1. The full grammar recognises the highlighted phrases in Example B.1 and assigns to them the structures given in the second column of Table B.1¹.

(B.1) *71 - letni pacjent z cukrzycą typu 2 rozpoznana 5 lat temu i nie leczoną przez ostatnie trzy lata został przyjęty do kliniki z powodu objawów klinicznych i biochemicznych niewyrównania cukrzycy. . . . W czasie hospitalizacji kontynuowano leczenie inhibitorem ACE zwiększając jego dawkę. . . . Kontrolne badanie lipidów za 4-6 tygodni. . . . Dieta cukrzycowa 1700 kcal., 5 posiłków/dobę.*

‘A **71 year old** patient with **diabetes type 2** diagnosed **5 years ago**, untreated over the last three years, was **hospitalised because of the clinical and biochemical parameters of unbalanced diabetes**. . . . During the hospitalisation, ACE inhibitor treatment was continued, the dose was increased. . . . Lipid control test for **4-6 weeks**. . . . **Diabetes diet 1700 kcal., 5 meals/day.**’

The simplified grammar recognises phrases highlighted in Example B.2.

(B.2) *71 - letni pacjent z cukrzycą typu 2 rozpoznana 5 lat temu i nie leczoną przez ostatnie **trzy lata** został przyjęty do kliniki z powodu objawów klinicznych i biochemicznych **niewyrównania** cukrzycy. . . . W czasie hospitalizacji kontynuowano leczenie inhibitorem ACE **zwiększając jego dawkę**. . . . Kontrolne badanie lipidów za **4-6 tygodni**. . . . **Dieta cukrzycowa 1700 kcal., 5 posiłków/dobę.***

‘A **71 year old** patient with diabetes **type 2** diagnosed **5 years ago**, untreated over the last **three years**, was hospitalised because of the clinical and biochemical parameters of **unbalanced diabetes**. . . . During the hospitalisation, ACE inhibitor treatment was continued, the **dose was increased**. . . . Lipid control test for **4-6 weeks**. . . . **Diabetes diet 1700 kcal., 5 meals/day.**’

¹ We use the following abbreviations: F_L is the FEATURE_L(IST) attribute, F and R are the FIRST and REST attributes of a list.

Table B.2 shows the structures assigned to the same text with the help of the simplified grammar. The grammar recognised two additional structures representing relative data and one piece of information on a therapy modification. As they are not related to diabetes, they are not taken into account in the final corpus annotation given in Table B.3. In this table, the second column contains the types of structures assigned to the texts indicated by the curly brackets, and the third column contains the structures assigned to the highlighted phrases within these texts.

Table B.1: Original grammar annotation

71 - letni	$\left[\begin{array}{l} id_p_str \\ ID_AGE \text{ "71"} \end{array} \right]$
pacjent z	
cukrzycą typu 2	$\left[\begin{array}{l} feature_l_str \\ \left[\begin{array}{l} feature_list \\ F \left[\begin{array}{l} d_type_str \\ D_TYPE \text{ second} \end{array} \right] \\ R \left[\begin{array}{l} feature_list \\ F \left[\begin{array}{l} diab_from_str \\ RELATIVE_DATA \left[\begin{array}{l} rel_data_str \\ D_NUM \text{ "5"} \\ D_UNIT \text{ u-year} \end{array} \right] \end{array} \right] \\ R \text{ null} \end{array} \right] \end{array} \right] \end{array} \right]$
rozpoznana 5 lat	
temu	
i nie leczoną przez ostatnie trzy lata został	
przyjęty do kliniki	$\left[\begin{array}{l} reaso_l_str \\ REASO_L \left[\begin{array}{l} reaso_list \\ FIRST \left[\begin{array}{l} d_contr_str \\ D_CONTROLL \text{ uncontrolled.t} \end{array} \right] \\ REST \text{ null} \end{array} \right] \end{array} \right]$
z powodu objawów	
klinicznych i bio- chemicznych niewyrów- nania cukrzycy	
...W czasie hospitalizacji kontynuowano leczenie in- hibitorem ACE zwiększając jego dawkę. ...Kontrolne badanie lipidów za 4-6 ty- godni. ...	
Dieta cukrzycowa 1700	$\left[\begin{array}{l} diet_str \\ DIET_TYPE \text{ d_diab.t} \\ CAL_MIN \text{ "1700"} \\ MEALS_MIN \text{ "5"} \end{array} \right]$
kcal., 5 posiłków	
/dobę.	

Table B.2: Simplified grammar annotation

71 -letni	$\left[\begin{array}{l} id_p_str \\ ID_AGE \text{ "71"} \end{array} \right]$
pacjent z cukrzycą	
typu 2	$\left[\begin{array}{l} d_type_str \\ D_TYPE \text{ second} \end{array} \right]$
rozpoznana	
5 lat temu	$\left[\begin{array}{l} rel_data_str \\ D_NUM \text{ "5"} \\ D_UNIT \text{ u_year} \end{array} \right]$
i nie leczoną przez ostatnie	
trzy lata	$\left[\begin{array}{l} rel_data_str \\ D_NUM \text{ "3"} \\ D_UNIT \text{ u_year} \end{array} \right]$
został przyjęty do kliniki z powodu objawów klinicznych i biochemicznych	
niewyrównania	$\left[\begin{array}{l} d_contr_str \\ D_CONTROLL \text{ uncontrolled.t} \end{array} \right]$
cukrzycy. . . W czasie hospitalizacji kontynuowano leczenie inhibitorem ACE	
zwiększając jego dawkę	$\left[\begin{array}{l} corr_str \\ DOSE_MODIFF \text{ yes} \end{array} \right]$
. . . . Kontrolne badanie lipidów za 4-	
6 tygodni	$\left[\begin{array}{l} rel_data_str \\ D_NUM \text{ "6"} \\ D_UNIT \text{ u_week} \end{array} \right]$
. . . .	
Dieta cukrzycowa	$\left[\begin{array}{l} diet_str \\ DIET_TYPE \text{ d_diab.t} \end{array} \right]$
1700 kcal.	$\left[\begin{array}{l} diet_str \\ CAL_MIN \text{ "1700"} \end{array} \right]$
,	
5 posiłków	$\left[\begin{array}{l} diet_str \\ MEALS_MIN \text{ "5"} \end{array} \right]$
/dobę.	

Table B.3: Result of comparison

71 -letni	<i>id_p_str</i>	[ID_AGE "71"]
pacjent z		
cukrzycą typu 2 rozpoznana 5 lat temu	} <i>feature_L_str</i>	[<i>d_type_str</i> D_TYPE <i>second</i>]
		[<i>rel_data_str</i> D_NUM "3" D_UNIT <i>u_year</i>]
i nie leczoną przez ostatnie trzy lata został		
przyjęty do kliniki z powodu objawów klinicznych i bio- chemicznych niewyrównania cukrzycy	} <i>reaso_L_str</i>	[<i>d_contr_str</i> D_CONTROLL <i>uncontrolled_t</i>]
. . . W czasie hospitalizacji kontynuowano leczenie inhibitorem ACE zwiększając jego dawkę. . . Kontrolne badanie lipidów za 4-6 tygodni. . .		
Dieta cukrzycowa 1700 kcal. , 5 posiłków	} <i>diet_str</i>	[DIET_TYPE <i>d_diab.t</i> [CAL_MIN "1700"]
		[MEALS_MIN "5"]
/dobę.		

C

Shallow grammar for phrase recognition

The following convention is used in the grammar:

- the upper case words represent grammar non-terminal symbols;
- non-terminal subscripts correspond to C – case, G – gender and N – number values, the same indexes in a rule represent unification of the values;
- the small caps names refer to POS tags and grammatical category values;
- IR refers to any POS;
- “|” denotes an alternative;
- “?” denotes optionality;
- N category refers to inflective nominal elements (*subst* and *ger*);
- NC category refers to elements that may be recognised as nominals but don’t inflect such as: foreign *foreign_subst*, *foreign* (foreign words), abbreviations of nominal phrases (*brev:pun:np*, *brev:pun:nphr*, *brev:npun:np*, *brev:npun:nphr*);
- AJ category refers to inflective adjective elements (*adj* and *ppas*);
- AC category refers to elements that may be recognised as adjectives but don’t inflect;
- NZ, AZ are words ignored during the terminology phrase construction;
- \hat{X} notation is used to mark cases in which morphological description of the resulting phrase should be copied from X-th element of the rule instead from the first one (e.g. case, gender and number of an adjective phrase consisting of an adverb and an adjective should be the same as those of the adjective);
- the first non-terminal in a rule is created from elements described by subsequent elements (terminals or non-terminals joined by “+”, it can be involved in the phrase recognition by the subsequent sets of rules.

Grammar

The first set of rules

N subst | ger
NC $\text{brev}_{npun,nw}$ | $\text{brev}_{npun,nw}$
NC $\text{brev}_{pun,nw}$ + “.”? | $\text{brev}_{pun,nw}$ + “.”?
NC $\text{brev}_{npun,nphr}$ | $\text{brev}_{npun,nphr}$
NC $\text{brev}_{pun,nphr}$ + “.”? | $\text{brev}_{pun,nphr}$ + “.”?
NC (foreign_subst | foreign) + foreign? + foreign?
AJ adj | ppas
AJ $\hat{2}$ adv? + ($\text{adj}_{C,G,N}$ | $\text{ppas}_{C,G,N}$)

AC brev_{adjw,npun} | brev_{adjw,pun} + “.”?
 CN “i”
 NZ subst(lemma=to/co/obrąb/kierunek/cel/czas/możliwość/...
 NZ subst(lemma=styczeń/luty/marzec/kwiecień/maj/czerwiec/...
 NZ subst(lemma=miesiąc/rok/tydzień
 AZ IR(lemma=aktualny/daleki/gdy/pewien/wzgląd/ten/inny/sam/
 niektóry/wczesny/wyznaczony/który...

The second set of rules

A AJ + adv?
 A³ AC + “-” + AJ_{C,G,N}
 A³ adja + “-” + AJ_{C,G,N}
 AJ² adv + AJ_{C,G,N}
 AC AC + “-” + AC
 AC AC + AC
 N AJ_{C,G,N} + N_{C,G,N}
 N N + NC

The third set of rules

ADJP A
 ADJP A_{C,G,N} + A_{C,G,N}? + A_{C,G,N}? + A_{C,G,N}? + CN + A_{C,G,N}
 ADJP A_{C,G,N} + A_{C,G,N}? + A_{C,G,N}? + A_{C,G,N}? + A_{C,G,N} + AC?
 ADJP² AC + A
 ADJP A AC
 N N + AC
 N N_{C,G,N} + AJ_{C,G,N}

The fourth set of rules

NB² NC + ADJP_C
 NB² AC + N
 NB N + AC
 NB ADJP_C + NC
 NB ADJP_{C,G,N}? + N_{C,G,N} + ADJP_{C,G,N}?

The fifth set of rules

NH NB_{case=nom/dat/acc/inst/loc} + NB_{gen}
 NH NB_{case=nom/dat/acc/inst/loc} + NB_{gen} + CN + NB_{gen}
 NH NB NC

The sixth set of rules

$NG\ NH + NB_{gen}$
 $NG\ (NH_{C,G,N} \mid NB_{C,G,N}) + ADJP_{C,G,N}?$

The seventh set of rules

$NX\ NG + NG_{gen}$
 $NX\ NG_C + \text{“,”?} + NG_C + CN + NG_C$
 $NX\ NG_C + CN + NG_C$
 $NX\ NG$

The eight set of rules

$X\ NX + NX_{gen}$
 $X\ NX + NC + \text{“.”}$
 $X\ NC + NX_{gen}$
 $X\ NC$
 $X\ NX_{C,G,N} + ADJP_{C,G,N}$
 $X\ NX$

D

Top 100 medical phrases

In the appendix, we give the 100 top terms extracted from the medical corpus containing discharge documents from the children’s hospital. The term extraction procedure used the C-value method where term candidates were selected with the help of the NPMI modification. The table contains terms in traditional base forms and their English translations, the number of occurrences as maximal phrases (full) and as nested phrases, the number of contexts counted as the maximum from different left and right words contexts counted separately, and C-values.

Table D.1: 100 top terms extracted from medical documents

Nb.	Term	Full	Nested	Cont.	C-value
1	<i>kod pacjenta</i> ‘patient code’	3,116	0	0	3,116.00
2	<i>wynik badania</i> ‘examination result’	1,754	1,391	34	3,104.09
3	<i>karta informacyjna leczenia szpitalnego</i> ‘hospital treatment information card’	1,280	1	1	2,560.00
4	<i>wzór białych krwinek</i> ‘pattern of white blood cells’	1,501	0	0	2,379.03
5	<i>układ kielichowo-miedniczkowy poszerzony</i> ‘widened pyelocalyceal system’	859	2	2	1,996.86
6	<i>stan ogólny dobry</i> ‘good general condition’	1,234	0	0	1,955.84
7	<i>stan ogólny</i> ‘general condition’	580	1,459	14	1,934.79
8	<i>wynik badań dodatkowych</i> ‘additional examinations results’	1,187	65	2	1,932.86
9	<i>jama brzuszna</i> ‘abdominal cavity’	958	882	17	1,788.12
10	<i>pęcherz moczowy</i> ‘bladder’	565	1,275	22	1,782.05
11	<i>oddział chirurgiczno-urazowy</i> ‘surgical and casualty ward’	833	0	0	1,666.00
12	<i>karta informacyjna</i> ‘information card’	546	1,588	3	1,604.67
13	<i>pęcherzyk żółciowy prawidłowy</i> ‘normal gallbladder’	991	1	1	1,570.70
14	<i>płytki krwi</i> ‘platelet’	1,531	23	5	1,549.40
15	<i>badanie ogólne</i> ‘general examination’	1,439	65	9	1,496.78
16	<i>nerka prawidłowej wielkości</i> ‘kidney of normal size’	943	0	0	1,494.62
17	<i>stan dobry</i> ‘good condition’	1,348	0	0	1,348.00

Table D.1: (continued)

Nb.	Term	Full	Nested	Cont.	C-value
18	<i>moczowód niewidoczny</i> ‘ureter invisible’	1,331	2	2	1,332.00
19	<i>USG jamy brzusznej</i> ‘ultrasound of the abdominal cavity’	744	78	11	1,291.60
20	<i>pęcherz moczowy wypełniony</i> ‘filled bladder’	775	35	3	1,265.33
21	<i>ciało ketonowe</i> ‘ketone body’	1,234	7	2	1,237.50
22	<i>prawidłowa echogeniczność</i> ‘normal echogenicity’	1,231	6	4	1,235.50
23	<i>ciężar ciała</i> ‘body weight’	1,226	0	0	1,226.00
24	<i>echogeniczność miąższu</i> ‘echogenicity of the parenchyma’	1,191	37	3	1,215.67
25	<i>prawidłowa wielkość</i> ‘normal size’	28	1,339	8	1,199.63
26	<i>oddział chirurgii</i> ‘surgical ward’	1,191	5	2	1,193.50
27	<i>badanie</i> ‘examination’	3,359	8,252	161	1,155.97
28	<i>zalecenie lekarskie</i> ‘medical recommendation’	1,154	0	0	1,154.00
29	<i>pęcherzyk żółciowy</i> ‘gallbladder’	118	1,043	7	1,012.00
30	<i>leczenie szpitalne</i> ‘hospital treatment’	1	1,284	4	964.00
31	<i>zastosowane leczenie</i> ‘applied treatment’	919	1	1	919.00
32	<i>oddział</i> ‘ward’	4,750	4,019	53	869.32
33	<i>fala ostra</i> ‘acute wave’	526	329	32	844.72
34	<i>wskaźnik protrombinowy</i> ‘prothrombin ratio’	825	1	1	825.00
35	<i>badanie przedmiotowe</i> ‘subjective examination’	715	38	4	743.50
36	<i>grupa krwi</i> ‘blood group’	697	2	2	698.00
37	<i>posiew moczu</i> ‘urine culture’	608	72	11	673.45
38	<i>składnik mineralny</i> ‘mineral component’	661	7	3	665.67
39	<i>fala wolna</i> ‘slow wave’	394	242	31	628.19
40	<i>nerka prawidłowej wielkości</i> ‘kidney of normal size’	383	2	2	608.63
41	<i>oddział rehabilitacji</i> ‘rehabilitation ward’	501	147	3	599.00
42	<i>badanie ogólne moczu</i> ‘general urine examination’	358	23	7	598.66
43	<i>test lateksowy</i> ‘latex test’	579	2	1	579.00
44	<i>stan</i> ‘condition’	908	4,990	36	575.94
45	<i>wsteczny odpływ pęcherzowo-moczowodowy</i> ‘vesicoureteral reflux’	240	6	5	568.41
46	<i>badanie laboratoryjne</i> ‘laboratory examination’	547	23	8	567.13
47	<i>oddział chirurgiczny</i> ‘surgical ward’	2	835	3	558.67
48	<i>badanie dodatkowe</i> ‘additional examination’	448	121	6	548.83
49	<i>klatka piersiowa</i> ‘chest’	404	128	26	527.08

Table D.1: (continued)

Nb.	Term	Full	Nested	Cont.	C-value
50	<i>bilirubina całkowita</i> ‘total bilirubin’	526	2	2	527.00
51	<i>ból głowy</i> ‘headache’	344	180	33	518.55
52	<i>ordynator oddziału</i> ‘head of the department’	499	1	1	499.00
53	<i>dno oka</i> ‘fundus of the eye’	495	1	1	495.00
54	<i>data urodzenia</i> ‘birth date’	486	0	0	486.00
55	<i>posiew ogólny</i> ‘general culture’	470	6	3	474.00
56	<i>zastosowane leczenie</i> ‘applied treatment’	464	1	1	464.00
57	<i>cystografia mikcyjna</i> ‘voiding cystourethrogram’	442	21	7	460.00
58	<i>mocz</i> ‘urine’	2,968	1,508	46	444.32
59	<i>ból brzucha</i> ‘abdominal pain’	398	48	16	443.00
60	<i>dieta lekkostrawna</i> ‘easily digested diet’	439	5	4	442.75
61	<i>leczenie diagnostyka zalecenia</i> ‘treatment diagnosis recommendation’	278	0	0	440.62
62	<i>krew</i> ‘blood’	1,444	3,013	37	437.56
63	<i>leczenie</i> ‘treatment’	3,332	962	47	427.35
64	<i>drogi oddechowe</i> ‘respiratory tract’	12	453	10	419.70
65	<i>szpital</i> ‘hospital’	3,750	316	24	405.28
66	<i>wynik</i> ‘result’	310	3,756	74	401.52
67	<i>cewka moczowa</i> ‘urethra’	269	137	19	398.79
68	<i>jama opłucnowa wolna</i> ‘no inflammation in pleural cavity’	247	2	2	393.07
69	<i>czynność podstawowa EEG zachowana</i> ‘normal EEG activity retained’	193	0	0	386.00
70	<i>zmiana ogniskowa</i> ‘lesion’	367	12	6	377.00
71	<i>zalecenie kontynuowania leczenia</i> ‘recommendation for treatment continuation’	223	3	2	355.82
72	<i>oddział neurologii</i> ‘neurology ward’	340	18	4	353.50
73	<i>antygen lamblii</i> ‘lamblia antigen’	347	12	2	353.00
74	<i>mg</i> ‘mg’	3,448	10	2	345.30
75	<i>opieka poradni</i> ‘clinic care’	26	329	30	344.03
76	<i>zapis nieprawidłowy</i> ‘abnormal record’	336	1	1	336.00
77	<i>infekcja dróg oddechowych</i> ‘infection of the respiratory tract’	153	63	15	335.70
78	<i>strona prawy</i> ‘right side’	319	17	6	333.17
79	<i>masa ciała</i> ‘body weight’	262	77	12	332.58
80	<i>pacjent</i> ‘patient’	233	3,224	23	331.68
81	<i>zapalenie płuc</i> ‘pneumonia’	226	111	19	331.16
82	<i>zalecenie</i> ‘recommendation’	2,715	573	73	328.02
83	<i>chemia kliniczny</i> ‘clinical chemistry’	325	0	0	325.00
84	<i>zakażenie układu moczowego</i>	188	18	9	323.33

Table D.1: (continued)

Nb.	Term	Full	Nested	Cont.	C-value
	‘urinary tract infection’				
85	<i>stan prawidłowy</i> ‘normal condition’	322	1	1	322.00
86	<i>infekcja górnych dróg oddechowych</i> ‘infection of the upper respiratory tract’	126	39	9	321.33
87	<i>opieka pediatryczna</i> ‘pediatric care’	317	6	2	320.00
88	<i>rozpoznanie</i> ‘diagnosis’	3,150	50	34	319.85
89	<i>układ kielichowo-miedniczkowy</i> ‘pyelocalyceal system’	117	45	13	317.08
90	<i>pleć</i> ‘sex’	3,117	41	6	315.12
91	<i>strona lewa</i> ‘left side’	298	19	10	315.10
92	<i>epikryza</i> ‘epicrisis’	3,108	6	5	311.28
93	<i>wykładnik stanu zapalnego</i> ‘inflammation exponent’	131	72	10	310.34
94	<i>stan ogólny średni</i> ‘average general condition’	194	1	1	307.48
95	<i>górne drogi oddechowe</i> ‘upper respiratory tract’	2	220	7	302.05
96	<i>fosfataza alkaliczna</i> ‘alkaline phosphatase’	297	6	2	300.00
97	<i>granica normy</i> ‘limit of the normal range’	270	33	10	299.70
98	<i>sen spontaniczny</i> ‘spontaneous sleep’	295	1	1	295.00
99	<i>MCV</i> ‘MCV’	2,938	0	0	293.80
100	<i>istotne odchylenie</i> ‘significant deviation’	290	0	0	290.00

References

- Ananiadou, S., Nenadic, G., Mima, H., and Tsujii, J. (2006). Mining Biomedical Terminology from Literature. In P. Hacken, editor, *Terminology, Computing and Translation*, pages 117–140. Gunter Narr Verlag, Tübingen.
- Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, **34**(4), 555–596.
- Bagga, A. and Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566. European Language Resources Association.
- Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Barrón-Cedeno, A., Sierra, G., Drouin, P., and Ananiadou, S. (2009). An Improved Automatic Term Recognition Method for Spanish. In *Computational Linguistics and Intelligent Text Processing*, volume 5449 of *Lecture Notes in Computer Science*, pages 125–136. Springer-Verlag, Berlin, Heidelberg.
- Basili, R., Moschitti, A., Pazienza, M., and Zanzotto, F. (2001). A Contrastive Approach to Term Extraction. In *Proceedings of the 4th Terminology and Artificial Intelligence Conference*, pages 119–128.
- Bański, P. and Przepiórkowski, P. (2009). Stand-off TEI Annotation: the Case of the National Corpus of Polish. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 64–67, Singapore.
- Bembenik, R., Skonieczny, Ł., Rybiński, H., Kryszkiewicz, M., and Niezgódka, M., editors (2013). *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*. Springer-Verlag, Cham, Heidelberg, New York, Dordrecht, London.
- Bąk, P. (1984). *Gramatyka języka polskiego*. Wiedza Powszechna, Warszawa.
- Boillat, P. and Kjaerum, M. (2014). *Handbook on European Data Protection Law*. European Union Agency for Fundamental Rights, Luxembourg.
- Borucki, B. (2009). Metodyka ochrony poufności i bezpieczeństwa medycznych danych osobowych. *Ultrasonografia*, **9**(36), 9–20.
- Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen. Gesellschaft für Sprachtechnologie & Computerlinguistik.
- Broda, B., Derwojedowa, M., and Piasecki, M. (2008). Recognition of Structured Collocations in an Inflective Language. *Systems Science*, **34**(4), 27–36.

- Buczyński, A. and Okniński, T. (2005). Program Kolokacje. <http://www.mimuw.edu.pl/polszczyzna/kolokacje/index-en.htm>.
- Buczyński, A. and Przepiórkowski, A. (2009). Spejd: A Shallow Processing and Morphological Disambiguation Tool. In Z. Vetulani and H. Uszkoreit, editors, *Human Language Technology: Challenges of the Information Society*, volume 5603 of *Lecture Notes in Artificial Intelligence*, pages 131–141. Springer-Verlag, Berlin, Heidelberg.
- Burnard, L. (2007). Reference Guide for the British National Corpus. <http://www.natcorp.ox.ac.uk/docs/URG/>.
- Cabré, M. T. (1999). *Terminology. Theory, Methods and Applications*. John Benjamins, Amsterdam, Philadelphia.
- Chang, J. T., Schütze, H., and Altman, R. B. (2002). Creating an Online Dictionary of Abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*, **9**, 612–620.
- Charniak, E. and Johnson, M. (2005). Coarse-to-fine n -best Parsing and Max-Ent Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Chiticariu, L., Li, Y., and Reiss, F. R. (2013). Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 827–832. Association for Computational Linguistics.
- Chung, T. M. and Nation, P. (2004). Identifying Technical Vocabulary. *System*, **32**(2), 251–263.
- Cooper, G. F. and Miller, R. A. (1998). An Experiment Comparing Lexical and Statistical Methods for Extracting MeSH Terms from Clinical Free Text. *JAMIA*, **5**(1), 62–75.
- Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*. The University of Sheffield. <http://gate.ac.uk/userguide>.
- Dalianis, H. and Velupillai, S. (2010). De-identifying Swedish Clinical Text — Refinement of a Gold Standard and Experiments with Conditional Random Fields. *Journal of Biomedical Semantics*, **1**(4), 1–6.
- Damerau, F. J. (1993). Generating and Evaluating Domain-oriented Multi-word Terms from Texts. *Information Processing and Management*, **29**(4), 433–447.
- Doroszewski, W., editor (1958-1969). *Słownik języka polskiego*. Państwowe Wydawnictwo Naukowe, Warszawa.
- Dridan, R. and Oepen, S. (2012). Tokenization: Returning to a Long Solved Problem. A Survey, Contrastive Experiment, Recommendations, and Toolkit. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, pages 378–382, Stroudsburg, PA, USA. Association for Computational Linguistics.

- Drozdzyński, W., Krieger, H.-U., Piskorski, J., Schäfer, U., and Xu, F. (2004). Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications. *German AI Journal KI-Zeitschrift*, **01/04**, 17–23.
- El Emam, K. and Arbuckle, L. (2013). *Anonymizing Health Data*. O’Reilly Media.
- Emam, K. E. E. (2013). *Guide to the De-Identification of Personal Health Information*. CRC Press, Taylor & Francis Group.
- Emele, M. C. (1994). TFS — The Typed Feature Structure Representation Formalism. In *Proceedings of the International Workshop on Shareable Natural Language Resources*.
- Erjavec, T., Tateisi, Y., Kim, J.-d., Ohta, T., and Tsujii, J. (2003). Encoding Biomedical Resources in TEI: the Case of the GENIA Corpus. In *Proceedings of the ACL 2003, Workshop on Natural Language Processing in Biomedicine*, pages 97–104. Association for Computational Linguistics.
- Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, **165**(1), 91–134.
- Fedorenko, D., Astrakhantsev, N., and Turdakov, D. (2013). Automatic Recognition of Domain-Specific Terms: An Experimental Evaluation. In N. Vassilieva, D. Turdakov, and V. Ivanov, editors, *SYRCoDIS*, volume 1031 of *CEUR Workshop Proceedings*, pages 15–23. CEUR-WS.org.
- Ferrández, O., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H., and Meystre, S. M. (2012). Evaluating Current Automatic De-Identification Methods with Veteran’s Health Administration Clinical Documents. *BMC Medical Research Methodology*, **12**(109), 16.
- Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *International Journal on Digital Libraries*, **3**, 115–130.
- Freitag, D. and McCallum, A. (2000). Information Extraction with HMM Structures Learned by Stochastic Optimization. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 584–589. Association for the Advancement of Artificial Intelligence.
- Friedlin, J. and McDonald, C. (2008). A Software Tool for Removing Patient Identifying Information from Clinical Documents. *Journal of the American Medical Informatics Association*, **15**(5), 601–610.
- Gardner, J. J. and Xiong, L. (2009). An Integrated Framework for De-identifying Unstructured Medical Data. *Data and Knowledge Engineering*, **68**(12), 1441–1451.
- Gelbukh, A., Sidorov, G., Lavin-Villa, E., and Chanona-Hernandez, L. (2010). Automatic Term Extraction Using Log-likelihood Based Comparison with General Reference Corpus. In *Proceedings of the Natural Language Processing and Information Systems, and 15th International Conference on Applications of Natural Language to Information Systems*, volume 6177 of *Lecture Notes in Computer Science*, pages 248–255. Springer-Verlag, Berlin, Heidelberg.

- Gęsicki, Ł. and Gęsicki, M. (1996). *Słownik terminów ekonomiczno-prawnych*. Interfart, Łódź.
- Gold, S., Elhadad, N., Zhu, X., Cimino, J. J., and Hripcsak, G. (2008). Extracting Structured Medication Event Information from Discharge Summaries. In *AMIA Annual Symposium Proceedings*, page 237–241. American Medical Informatics Association.
- Głowiński, M., Kostkiewiczowa, T., and Okopień-Sławińska, A., editors (2010). *Słownik terminów literackich*. Ossolineum.
- Graliński, F., Jassem, K., Marcińczuk, M., and Wawrzyniak, P. (2009). Named Entity Recognition in Machine Anonymization. In M. A. Kłopotek, A. Przepiórkowski, A. T. Wierzchoń, and K. Trojanowski, editors, *Recent Advances in Intelligent Information Systems*, pages 247–260. Akademicka Oficyna Wydawnicza Exit, Warszawa.
- Grigonytė, G., Rimkutė, E., Utkā, A., and Boizou, L. (2011). Experiments on Lithuanian Term Extraction. In *Proceedings of the 18th Nordic Conference of Computational Linguistics*, pages 82–89. Northern European Association for Language Technology.
- Grishman, R. and Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Grouin, C. and Névėol, A. (2014). De-identification of Clinical Notes in French: Towards a Protocol for Reference Corpus Development. *Journal of Biomedical Informatics*, **50**, 151–161.
- Grucza, F. (1991a). Teoretyczne podstawy terminologii. In Grucza (1991b), chapter Terminologia — jej przedmiot, status i znaczenie, pages 11–43.
- Grucza, F., editor (1991b). *Teoretyczne podstawy terminologii*. Zakład Narodowy imienia Ossolińskich, Wrocław, Warszawa, Kraków.
- Gupta, D., Saul, M., and Gilbertson, J. (2004). Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research. *American Journal of Clinical Pathology*, **121**(1), 176–186.
- Gupta, S. and Manning, C. (2014a). SPIED: Stanford Pattern Based Information Extraction and Diagnostics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 38–44, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Gupta, S. and Manning, C. D. (2014b). Improved Pattern Learning for Bootstrapped Entity Extraction. In R. Morante and W. Yih, editors, *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 98–108. Association for Computational Linguistics.
- Hajnicz, E. (2011). Najbardziej znane korpusy tekstów. Opracowanie przeglądowe. Raport techniczny 1022, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa.
- Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E. A. (2003). Automatic Document Metadata Extraction Using Support Vector Machines.

- In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 37–48. Institute of Electrical and Electronics Engineers.
- Hara, K. (2006). Applying a SVM Based Chunker and a Text Classifier to the Deid Challenge. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.
- Hasegawa, T., Sekine, S., and Grishman, R. (2004). Discovering Relations among Named Entities from Large Corpora. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, pages 415–422, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hassler, M. and Fliedl, G. (2006). Text Preparation through Extended Tokenization. *Data Mining VII: Data, Text and Web Mining and their Business Applications*, **37**, 13–21.
- Hisamitsu, T. and Tsujii, J. (2003). Measuring Term Representativeness. In M. T. Pazienza, editor, *Information Extraction in the Web Era: Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents*, volume 2700 of *Lecture Notes in Computer Science*, pages 45–76. Springer-Verlag, Berlin, Heidelberg.
- Hobbs, J. R. and Riloff, E. (2010). Information Extraction. In N. Indurkha and F. J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 511–532. CRC Press, Taylor and Francis Group, Boca Raton, Florida.
- Hoste, V., Vanopstal, K., Lefever, E., and Delaere, I. (2010). Classification-based Scientific Term Detection in Patient Information. *Terminology*, **16**(1), 1–29.
- Hoste, V., Vanopstal, K., Lefever, E., and Delaere, I. (2011). Automatic Extraction of Semantic Relations between Medical Entities: A Rule Based Approach. *Journal of Biomedical Semantics*, **2**.
- Ittoo, A. and Bouma, G. (2013). Term Extraction from Sparse, Ungrammatical Domain-Specific Documents. *Expert Systems with Applications*, **40**(7), 2530–2540.
- Ji, L., Sum, M., Lu, Q., Li, W., and Chen, Y.-R. (2007). Chinese Terminology Extraction Using Window-Based Contextual Information. In A. F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18–24, 2007*, volume 4394 of *Lecture Notes in Computer Science*, pages 62–74. Springer-Verlag, Berlin, Heidelberg.
- Kageura, K. and Umino, B. (1996). Method for Automatic Term Recognition. A Review. *Terminology*, **3:2**, 259–289.
- Karwańska, D. and Przepiórkowski, A. (2009). On the Evaluation of Two Polish Taggers. In *The proceedings of Practical Applications in Language and Computers PALC 2009*.
- Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA Corpus – a Semantically Annotated Corpus for Bio-Textmining. *Bioinformatics*, **19**(suppl 1), 180–182.
- Kim, J.-D., Ohta, T., and Tsujii, J. (2008). Corpus Annotation for Mining Biomedical Events from Literature. *BMC Bioinformatics*, **9**(1), 10.

- Kluegl, P., Atzmueller, M., and Puppe, F. (2008). Test-Driven Development of Complex Information Extraction Systems using TextMarker. In G. J. Naplepa and J. Baumeister, editors, *4th International Workshop on Knowledge Engineering and Software Engineering, 31th German Conference on Artificial Intelligence*, pages 19–30.
- Knoth, P., Schmidt, M., Smrž, P., and Zdráhal, Z. (2009). Towards a Framework for Comparing Automatic Term Recognition Methods. In *Znalosti 2009*, pages 83–94. Faculty of Informatics and Information Technology STU.
- Kobyliński, L. (2012). Mining Class Association Rules for Word Sense Disambiguation. In P. Bouvry, M. A. Kłopotek, F. Lerepovost, M. Marciniak, A. Mykowiecka, and H. Rybiński, editors, *Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, pages 307–318. Springer-Verlag, Berlin, Heidelberg.
- Koeva, S. (2007). Multi-Word Term Extraction for Bulgarian. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Kokkinakis, D. and Thurin, A. (2008). Applying MeSH® to the (Swedish) Clinical Domain - Evaluation and Lessons Learned. In *Proceedings of the 6th Scandinavian Health Informatics and the 12th Swedish National Term Conference*, pages 37–41.
- Korkontzelos, I., Klapaftis, I. P., and Manandhar, S. (2008). Reviewing and Evaluating Automatic Term Recognition Techniques. In *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Artificial Intelligence*, pages 248–259. Springer-Verlag, Berlin, Heidelberg.
- Koukourikos, A., Karampiperis, P., Vouros, G. A., and Karkaletsis, V. (2012). Using Open Information Extraction and Linked Open Data towards Ontology Enrichment and Alignment. In M. Bajec and J. Eder, editors, *CAiSE Workshops*, volume 112 of *Lecture Notes in Business Information Processing*, pages 117–122. Springer-Verlag, Berlin, Heidelberg.
- Kozakiewicz, S., editor (2014). *Słownik terminologiczny sztuk pięknych*. Państwowe Wydawnictwo Naukowe, Warszawa.
- Kristjansson, T., Culotta, A., Viola, P., and McCallum, A. (2004). Interactive Information Extraction with Constrained Conditional Random Fields. In *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 412–418. Association for the Advancement of Artificial Intelligence.
- Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Lüdeling, A. and Kytö, M., editors (2008). *Corpus Linguistics. An International Handbook*. Mouton de Gruyter.
- Li, Y., Bontcheva, K., and Cunningham, H. (2005). SVM Based Learning System for Information Extraction. In *Proceedings of Sheffield Machine Learn-*

- ing Workshop*, volume 2266 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, Heidelberg.
- Lin, T., Mausam, and Etzioni, O. (2010). Commonsense from the Web: Relation Properties. In *AAAI Fall Symposium: Commonsense Knowledge*. Association for the Advancement of Artificial Intelligence.
- Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014). Integration of Linguistic and Web Information to Improve Biomedical Terminology Extraction. In *18th International Database Engineering & Applications Symposium*, pages 265–269. Association for Computing Machinery.
- Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.
- Marcińczuk, M., Kocóń, J., and Janicki, M. (2013). Liner2 – A Customizable Framework for Proper Names Recognition for Polish. In Bembenik *et al.* (2013), pages 231–253.
- Marciniak, M., editor (2010a). *Anotowany korpus dialogów telefonicznych*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.
- Marciniak, M. (2010b). Wyodrębnianie prostych fraz. In Marciniak (2010a), pages 217–230.
- Marciniak, M. and Mykowiecka, A. (2007). Automatic Processing of Diabetic Patients’ Hospital Documentation. In J. Piskorski, B. Pouliquen, R. Steinberger, and H. Tanev, editors, *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing at ACL 2007*, pages 35–42, Prague.
- Marciniak, M. and Mykowiecka, A. (2011a). Construction of a Medical Corpus Based on Information Extraction Results. *Control & Cybernetics*, **40**(2), 337–360.
- Marciniak, M. and Mykowiecka, A. (2011b). Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish. In *Proceedings of BioNLP 2011*, pages 92–100.
- Marciniak, M. and Mykowiecka, A. (2013). Terminology Extraction from Domain Texts in Polish. In Bembenik *et al.* (2013), pages 171–185.
- Marciniak, M. and Mykowiecka, A. (2014a). NPMI Driven Recognition of Nested Terms. In *Proceedings of the 4th International Workshop on Computational Terminology*, pages 33–41. Association for Computational Linguistics and Dublin City University.
- Marciniak, M. and Mykowiecka, A. (2014b). Terminology Extraction from Medical Texts in Polish. *Journal of Biomedical Semantics*, **5**.
- Marciniak, M. and Mykowiecka, A. (2015). Nested Term Recognition Driven by Word Connection Strength. *Terminology*, **21**(2).
- Marciniak, M., Mykowiecka, A., and Waszczuk, J. (2008). Automatyczne wypełnianie bazy danych pacjentów diabetologicznych na podstawie wypisów szpitalnych. In *Proceedings of INFOBAZY 2008*.
- Marciniak, M., Mykowiecka, A., and Rychlik, P. (2010). Medical Text Data Anonymization. *Journal of Medical Informatics & Technologies*, **16**, 83–88.

- Markó, K., Daumke, P., Schulz, S., and Hahn, U. (2003). Cross-language MeSH Indexing Using Morpho-Semantic Normalization. *AMIA 2003 Symposium Proceedings*, pages 425–429.
- Masarie, F. E. and Miller, R. A. (1987). Medical Subject Headings and Medical Terminology: An Analysis of Terminology Used in Hospital Charts. *Bulletin of the Medical Library Association*, **2**(75), 89–94.
- McCallum, A. and Li, W. (2003). Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Mcdonald, R. and Pereira, F. (2005). Identifying Gene and Protein Mentions in Text Using Conditional Random Fields. *BMC Bioinformatics*, **22**.
- Meystre, S., Friedlin, F., South, B., Shen, S., and Samore, M. (2010). Automatic De-Identification of Textual Documents in the Electronic Health Record: a Review of Recent Research. *BMC medical research methodology*, **10**(1), 70.
- Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., and Hurdle, J. F. (2008). Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *IMIA Yearbook 2008: Access to Health Information*, **2008**(1), 128–144.
- Milios, E., Zhang, Y., He, B., and Dong, L. (2003). Automatic Term Extraction and Document Similarity in Special Text Corpora. In *6th Conference of the Pacific Association for Computational Linguistics*, pages 275–284, Halifax, Nova Scotia, Canada.
- Mykowiecka, A. (2010). Anotacja pojęciami dziedzinowymi. In Marciniak (2010a), pages 83–98.
- Mykowiecka, A. and Marciniak, M. (2011a). Automatic Semantic Labeling of Medical Texts with Feature Structures. In I. Habernal and V. Matoušek, editors, *Text, Speech and Dialogue: 14th International Conference*, volume 6836 of *Lecture Notes in Artificial Intelligence*, pages 49–56, Berlin, Heidelberg. Springer-Verlag.
- Mykowiecka, A. and Marciniak, M. (2011b). Some Remarks on Automatic Semantic Annotation of a Medical Corpus. In *Proceedings of Third Louhi Workshop on Health Documentation Text Mining and Information Analysis at AIME*.
- Mykowiecka, A. and Marciniak, M. (2012a). Clustering of Medical Terms Based on Morpho-Syntactic Features. In *Proceedings of International Conference on Knowledge Engineering and Ontology Development*, pages 214–219. SciTePress.
- Mykowiecka, A. and Marciniak, M. (2012b). Combining Wordnet and Morphosyntactic Information in Terminology Clustering. In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai.
- Mykowiecka, A. and Marciniak, M. (2014). Attribute Value Acquisition through Clustering of Adjectives. In A. Przepiórkowski and M. Ogrodniczuk, editors, *Advances in Natural Language Processing: Proceedings of the 9th International*

- Conference on NLP*, volume 8686 of *Lecture Notes in Artificial Intelligence*, pages 92–104. Springer-Verlag, Berlin, Heidelberg.
- Mykowiecka, A. and Waszczuk, J. (2009). Semantic Annotation of City Transportation Information Dialogues Using CRF Method. In V. Matoušek and P. Mautner, editors, *Text, Speech and Dialogue: 13th International Conference*, volume 5729 of *Lecture Notes in Artificial Intelligence*, pages 411–419, Berlin, Heidelberg. Springer-Verlag.
- Mykowiecka, A., Marciniak, M., and Kupść, A. (2009). Rule-based Information Extraction from Patients’ Clinical Data. *Journal of Biomedical Informatics*, **42**, 923–936.
- Nakagawa, H. and Mori, T. (2003). Automatic Term Recognition Based on Statistics of Compound Nouns and their Components. *Terminology*, pages 201–219.
- Neamatullah, I., Douglass, M. M., Lehman, L.-W. H., Reisner, A. T., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., and Clifford, G. D. (2008). Automated De-Identification of Free-Text Medical Records. *BMC Medical Informatics and Decision Making*, **8**, 32.
- Nenadić, G., Spasić, I., and Ananiadou, S. (2002). Automatic Acronym Acquisition and Term Variation Management within Domain-specific Texts. In *In Bibliography 83 Proceedings of the 3rd International Conference on Language, Resources, and Evaluation (LREC-3)*, pages 2155–2162.
- Nenadic, G., Ananiadou, S., and McNaught, J. (2004). Enhancing Automatic Term Recognition through Recognition of Variation. In *Proceedings of Coling 2004*, pages 604–610, Geneva, Switzerland. COLING.
- Ohta, T., Pyysalo, S., Kim, J.-D., and Tsujii, J. (2010). A Re-Evaluation of Biomedical Named Entity-Term Relations. *J. Bioinformatics and Computational Biology*, **8**(5), 917–928.
- Okazaki, N. (2007). CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs). <http://www.chokkan.org/software/crfsuite/>.
- Okazaki, N. and Ananiadou, S. (2006). Building an Abbreviation Dictionary Using a Term Recognition Approach. *Bioinformatics*, **22**(24), 3089–3095.
- Onken, M., Riesmeier, J., Engel, M., Yabanci, A., Zabel, B., and Després, S. (2009). Reversible Anonymization of DICOM Images using Automatically Generated Policies. In K.-P. Adlassnig, B. Blobel, J. Mantas, and I. Masic, editors, *Medical Informatics in a United and Healthy Europe, Proceedings of MIE 2009*, pages 861–865. European Federation of Medical Informatics, IOS Press.
- Pantel, P. and Lin, D. (2001). A Statistical Corpus-Based Term Extractor. In *Advances in Artificial Intelligence*, volume 2056 of *Lecture Notes in Artificial Intelligence*, pages 36–46. Springer-Verlag, Berlin, Heidelberg.
- Park, Y. and Byrd, R. J. (2001). Hybrid Text Mining for Finding Abbreviations and their Definitions. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 126–133. Association for Computational Linguistics.

- Park, Y., Byrd, R. J., and Boguraev, B. K. (2002). Automatic Glossary Extraction: Beyond Terminology Identification. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pazienza, M., Pennacchiotti, M., and Zanzotto, F. (2005). Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. In S. Sirmakessis, editor, *Knowledge Mining. Proceedings of the NEMIS 2004 Final Conference*, volume 185 of *Studies in Fuzziness and Soft Computing*, pages 255–279. Springer-Verlag, Berlin, Heidelberg.
- Pecina, P. and Schlesinger, P. (2006). Combining Association Measures for Collocation Extraction. In N. Calzolari, C. Cardie, and P. Isabelle, editors, *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.
- Peng, F. and McCallum, A. (2004). Accurate Information Extraction from Research Papers using Conditional Random Fields. In D. M. Susan Dumais and S. Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 329–336, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Pereira, S., Neveol, A., Kerdelhué, G., Serrot, E., Joubert, M., and Darmoni, S. J. (2008). Using Multi-terminology Indexing for the Assignment of MeSH Descriptors to Health Resources in a French Online Catalogue. *AMIA 2008 Annual Symposium Proceedings*, pages 586–590.
- Pęzik, P. (2012). Wyszukiwarka PELCRA dla danych NKJP. In *Przepiórkowski et al. (2012)*, pages 253–273.
- Piasecki, M. (2007). Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *Task Quarterly*, **11**(1–2), 151–167.
- Piasecki, M. and Radziszewski, A. (2007). Polish Morphological Guesser Based on a Statistical A Tergo Index. In *Proceedings of the International Multiconference on Computer Science and Information Technology — 2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'07)*, pages 247–256.
- Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.
- Piskorski, J. (2008). ExPRESS: Extraction Pattern Recognition Engine and Specification Suite. In T. Hanneforth and K.-M. Würzner, editors, *Finite-state methods and natural language processing: 6th International Workshop, FSMNLP 2007*, pages 166–183. Universitätsverlag Potsdam.
- Polański, E. and Nowak, T. (2011). *Najnowszy podręcznik gramatyki języka polskiego*. PETRUS, Kraków.
- Poon, H. and Domingos, P. (2010). Unsupervised Ontology Induction from Text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 296–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

- Przepiórkowski, A. (2004). *Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary Version*. IPI PAN, Warszawa.
- Przepiórkowski, A. (2005). The IPI PAN Corpus in Numbers. In Z. Vetulani, editor, *Proceedings of the 2nd Language & Technology Conference*, Poznań, Poland.
- Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.
- Przepiórkowski, A., Kupść, A., Marciniak, M., and Mykowiecka, A. (2002). *Formalny opis języka polskiego: Teoria i implementacja*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.
- Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.
- Pęzik, P. (2014). Graph-based Analysis of Collocational Profiles. In V. Jesenšek and P. Grzybek, editors, *Phraseologie im Wörterbuch und Korpus*, pages 227–243. Zora 97, Maribor, Bielsko-Biała, Budapest, Kansas, Praha.
- Riloff, E. (1996a). Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2, AAAI'96*, pages 1044–1049. Association for the Advancement of Artificial Intelligence.
- Riloff, E. (1996b). An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. *Artificial Intelligence*, **85**, 101–134.
- Robertson, S. (2004). Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation*, **60**(5), 503–520.
- Rybicka-Nowacka, H. (1991). Teoretyczne podstawy terminologii. In Grucza (1991b), chapter Normalizacja polskiej terminologii technicznej, pages 141–157.
- Sager, J. C. (1990). *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam, Philadelphia.
- Saloni, Z. and Świdziński, M. (2001). *Składnia współczesnego języka polskiego*. Państwowe Wydawnictwo Naukowe, Warszawa.
- Salton, G. (1988). Syntactic Approaches to Automatic Book Indexing. In *Proceedings of the 26th Annual Meeting on Association for Computational Linguistics, ACL '88*, pages 204–210, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Santorini, B. (1990). Part-Of-Speech Tagging Guidelines for the Penn Treebank Project (3rd revision, 2nd printing). Technical Report MS-CIS-90-4, Department of Linguistics, University of Pennsylvania, Philadelphia, PA, USA.
- Savary, A. and Waszczuk, J. (2012). Narzędzia do anotacji jednostek nazewniczych. In Przepiórkowski *et al.* (2012), pages 225–252.
- Savary, A., Zaborowski, B., Krawczyk-Wieczorek, A., and Makowiecki, F. (2012). SEJFEK – a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. In *Proceedings of Cognitive Aspects of the Lexicon (COGALEX-III), a Workshop at COLING 2012*, Mumbai, India.

- Savova, G. K., Harris, M., Johnson, T., Pakhomov, S. V., and Chute, C. G. (2003). A Data-Driven Approach for Extracting “the Most Specific Term” for Ontology Development. *AMIA 2003 Annual Symposium Proceedings*, pages 579–583.
- Schoenmackers, S., Etzioni, O., Weld, D. S., and Davis, J. (2010). Learning First-Order Horn Clauses from Web Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1088–1098, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Seymore, K., McCallum, A., and Rosenfeld, R. (1999). Learning Hidden Markov Model Structure for Information Extraction. In *In AAAI 99 Workshop on Machine Learning for Information Extraction*, pages 37–42. Association for the Advancement of Artificial Intelligence.
- Sha, F. and Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 134–141, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Sperberg-McQueen, C. M. and Burnard, L. (2008). TEI P5: Guidelines for Electronic Text Encoding and Interchange. <http://www.tei-c.org/Vault/P5/1.3.0/doc/tei-p5-doc/en/html/>.
- Su, J., Yang, X., Hong, H., Tateisi, Y., and Tsujii, J. (2008). Coreference Resolution in Biomedical Texts: A Machine Learning Approach. In M. Ashburner, U. Leser, and D. Rebholz-Schuhmann, editors, *Ontologies and Text Mining for Life Sciences: Current Status and Future Perspectives*, number 08131 in Dagstuhl Seminar Proceedings. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany.
- Świdziński, M. and Woliński, M. (2009). A New Formal Definition of Polish Nominal Phrases. In M. Marciniak and A. Mykowiecka, editors, *Aspects of Natural Language Processing*, volume 5070 of *Information Systems and Applications*, pages 143–162. Springer-Verlag, Berlin, Heidelberg.
- Szarvas, G., Farkas, R., and Busa-Fekete, R. (2007). State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. *Journal of the American Medical Informatics Association*, **14**(5), 574–580.
- Szober, S. (1953). *Gramatyka języka polskiego*. Nasza Księgarnia, Warszawa.
- Tateisi, Y. and Tsujii, J. (2006). GENIA Annotation Guidelines for Treebanking. Technical Report TR-NLP-UT-2006-5, Tsujii Laboratory, University of Tokyo.
- Tateisi, Y. and Tsujii, J. (2006). GENIA Annotation Guidelines for Tokenization and POS Tagging. Technical Report TR-NLP-UT-2006-4, Tsujii Laboratory, University of Tokyo.
- Tomanek, K., Daumke, P., Enders, F., Huber, J., Theres, K., and Müller, M. (2012). An Interactive De-Identification-System. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine*, pages 97–104.
- Uzuner, Ö., Luo, Y., and Szolovits, P. (2007). Evaluating the State-of-the-art in Automatic De-identification. *Journal of the American Medical Informatics Association*, **14**(5), 550–563.

- Øvsthus, K., Innselset, K., M., B., and Kristiansen, M. (2005). Developing Automatic Term Extraction. Automatic Domain Specific Term Extraction for Norwegian. In *Proceedings of Terminology and Knowledge Engineering*, pages 50–62.
- Waszczuk, J. (2010). Anotacja pojęciami z wykorzystaniem metod maszynowego uczenia. In Marciniak (2010a), pages 137–149.
- Wermter, J. and Hahn, U. (2005). Massive Biomedical Term Discovery. In *Discovery Science*, volume 3735 of *Lecture Notes in Computer Science*, pages 281–293. Springer-Verlag, Berlin, Heidelberg.
- Wojnicki, S. (1991). Teoretyczne podstawy terminologii. In Grucza (1991b), chapter Subjęzyki specjalistyczne, pages 61–116.
- Woliński, M. (2006). Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Processing and Web Mining*, volume 36 of *Advances in Soft Computing*, pages 503–512. Springer-Verlag, Berlin.
- Woliński, M. (2014). Morfeusz Reloaded. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. European Language Resources Association.
- Wynne, M. (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxbow Books, Oxford.
- Xiao, R. (2008). Well-known and Influential Corpora: A Survey. In Lüdeling and Kytö (2008), pages 383–456.
- Xiao, R. (2010). Corpus Creation. In N. Indurkha and F. J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 147–166. CRC Press, Taylor and Francis Group, Boca Raton, FL.
- Yang, Y., Lu, Q., and Zhao, T. (2008). Chinese Term Extraction Using Minimal Resources. In D. Scott and H. Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 1033–1040.
- Zaborowski, B. (2012). Spejd 1.3.6 - User Manual. <http://zil.ipipan.waw.pl/Spejd/>.
- Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*, pages 2108–2113, Marrakech, Morocco. European Language Resources Association.

Index

- κ coefficient, 16, 76
- abbreviation, 34, 40
 - recognition, 90–91
- accuracy, 17
- anonymisation, 23–32
- ATR, *see* automatic term recognition
- automatic term recognition, *see* terminology extraction
- British National Corpus, 20
- C-value, 92–93, 106–108
- C/NC-value, 93, 112
- Conditional Random Fields, 9–10, 78–82
- context, 108–109
- coreference resolution, 7
- corpus, 19–21
- CRF, *see* Conditional Random Fields
- database, 69–71
- de-identification, *see* anonymisation
- domain corpus, 21
- F_β -score, 17
- F-score, 16–18
- GATE, 12
- gazetteer, 9, 12, 15, 53, 58–59
- general corpus, 20, 129
- GENIA, 21
- Guesser, 38
- hash file, 78–79, 102
- head element, 11, 95, 105, 134
- Hidden Markov Model, 9
- IDF, *see* inverse document frequency
- IE, *see* information extraction
- information extraction, 7, 9–12
 - machine learning, 9–10, 77–82
 - rule based, 8, 9, 54–69
 - unsupervised, 11–12
- information retrieval, 7, 88
- inverse document frequency, 89
- inverse word frequency, 96
- IR, *see* information retrieval
- log-likelihood, 94
- Medical Subject Headings, 116–118
- MeSH, *see* Medical Subject Headings
- Message Understanding Conference, 9
- morphological analyses, 38–39
- MUC, *see* Message Understanding Conference
- multi-hierarchy, 15, 55, 58, 65
- mutual information, 94
- named entity recognition, 7, 10, 26, 29
- Narodowy Korpus Języka Polskiego, 20, 34, 50, 129
- NER, *see* named entity recognition
- nested phrase, 111, 118–119
- nested term, 93, 99, 104
- NKJP, *see* Narodowy Korpus Języka Polskiego
- Normalised Pointwise Mutual Information, 118, 120–121
- NPMI, *see* Normalised Pointwise Mutual Information
- open information extraction, 12
- pointwise mutual information, 12
- Polish National Corpus, *see* Narodowy Korpus Języka Polskiego
- Polish tagset, 39, 141–143
- precision, 16–18
- recall, 16–18
- reversible anonymisation, 26, 30–31

- reversible coding, *see* reversible
anonymisation
- shallow grammar, 86, 102–104, 147–149
- shallow parsing, 10, 13
- simplified base form, 104–106
 - conversion, 133–136
- Spejd, 13–14
 - grammar, 47–49
- SProUT, 12, 14–15, 55–66
 - grammar, 60–66
 - tokenisation, 36–38
- stop-list, 87, 103
- Support Vector Machine, 9

- TaKIPI, 38–39
 - tokenisation, 36–38
- TEI guidelines, 50
- term, 84–85
 - normalisation, 90–92
 - variant, 89–90
- term frequency, 89
- termhood, 87, 96
- terminology, 83–84
- terminology extraction, 7, 100–109
 - process, 86–88
- TF, *see* term frequency
- TF-IDF, 88–89
- TFS, *see* typed feature structure
- tokenisation, 34–38, 46–47
- truncated phrase, 110, 118, 127, 128
- type hierarchy, 56
- typed feature structure, 14, 55–58

- UMLS, *see* Unified Medical Language
System
- unification, 12, 14, 15, 56, 62, 65–66, 69
- Unified Medical Language System, 27,
116
- unithood, 87, 96, 98, 118



KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI



UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY



The Project is co-financed by the European Union from resources of the European Social Fund

ISBN 978-83-63159-12-2

e-ISBN 978-83-63159-13-9