How specialised should a general dictionary be?

Špela Vintar

Abstract
The article discusses theoretical and methodological issues related to specialised vocabulary in the *Dictionary of Modern Slovene Language*. We address key questions such as the role of terminology in a general dictionary, user requirements and needs, the complexity of distinction between general and specialised terms, and finally corpus composition and corpus representativity. We propose a model where lexical items are categorised into three levels of termhood, and each level of specialisation requires a different strategy of lexicographical description. By illustrating possible relations between the proposed categories and the corpus-based methodology of candidate extraction we establish a working methodology for handling specialised units in a general dictionary.

Keywords: specialised vocabulary, general dictionary, terminology extraction, user requirements, specialised vs. general

1 INTRODUCTION

We all use specialised lexical items in our everyday lives, for the simple reason that nearly everyone engages in activities or fields which are not shared by all speakers of the language, and which involve the communication of specialised skills or know-how. It seems that as native speakers of a language we are equipped with an intuitive gauge of termhood by distinguishing between highly and less specialised items, and we often justify such intuitions with statements such as "This is sailors' jargon" or "I can't understand this medical gibberish". But would we expect to find such items in a general dictionary, or would we consult a dictionary at all when encountering them?

In this paper we present a series of reflections on the role of specialised vocabulary in a general dictionary, specifically the *Dictionary of Modern Slovene Language* (DMSL), considering various aspects from the established traditions and practices in Slovene lexicography, user requirements and profiles, corpus composition and representativeness, to lexicographic description and data presentation for different target groups. The aim of our discussion is to establish a methodological framework which would provide guidelines on the treatment of specialised vocabulary through all stages of dictionary creation, and which would be sustainable both in terms of adaptability and scalability to different target groups and in terms of labour intensity by employing (semi-)automatic techniques of data acquisition.

Clearly the above goal is not an easy one, and perhaps one would expect that such fundamental methodological questions have been extensively dealt with by lexicographers in related dictionary projects worldwide, and that their findings could easily be transferred into the Slovene language community. Surprisingly though, the body of literature with in-depth descriptions of methodological decisions regarding specialised lexis in general dictionaries is relatively lean, especially in comparison to the many studies dealing with specialised dictionaries, terminology or so-called LSP (language for special purposes); it is therefore necessary to draw conclusions from general dictionaries themselves, or occasionally their introductions. Moreover, the experiences and methodologies from related dictionary projects elsewhere are not directly replicable in the Slovene situation, firstly because of the strong influence of the specific lexicographic history in Slovenia, and secondly because of the currently prevailing social norms reflected in the official language policy. Both of these factors will inevitably influence the expectations of potential dictionary users and, as a consequence, the range of functions that the new dictionary should fulfil. The proposed methodological framework therefore relies on existing practices only as the point of departure from which an iterative cycle of improvements should evolve.

2 THE ROLE OF SPECIALISED VOCABULARY IN A GENERAL DICTIONARY

The tendency to include technical terms into general mono- and multilingual dictionaries has been on the increase since the 19th century (Boulanger 1996: 141), partly because the impact of science and technology on everyday life has been growing since the Enlightenment, but also due to the rising level of education and the inclusion of the so-called "technolects" into vernacular language use. The second part of the 19th century was a crucial period for the formation of basic terminology in Slovene, both in natural and human sciences, driven largely by numerous translations of scientific and reference works from German and other languages into Slovene (Prunč 2009).

From the 20th century onwards general language dictionaries gradually diminished their normative character and increasingly started to consider the expectations of users, which entailed the demand for a broad coverage of specialised items in a comprehensive dictionary. Landau (2001) even claims that contemporary comprehensive dictionaries seem as if multiple LSP dictionaries have been added to the traditional general language dictionary, mostly because emerging disciplines continually produce more new lexical items than general language. The reasons for the growing ratio of terms in general dictionaries are summarized by Josselin-Leray (2005) as follows:

a) An almost two centuries long tradition in lexicography of increasingly including specialised lexis in general dictionaries.

b) The growing trend of despecialisation (determinologisation), the process through which specialised terms move into everyday language and typically modify or broaden their meaning.

c) The didactic role of general dictionaries, which through working to meet the needs of EFL learners revolutionised English lexicography and brought profound changes to the dictionary-making process worldwide.

d) Striving for comprehensiveness, whereby a single reference work aims to satisfy the needs of the broadest possible target audience.

e) The expectations and requirements of users, who are today better informed and more interested in science and technology than in the past.

All of these reasons apply to the Slovene language community, and thus build a case for a strong representation of terminology in the new contemporary dictionary of Slovene.

The only existing comprehensive dictionary of Slovene, the *Dictionary of Slovene Literary Language* (DSLL), gives terminology an important role – indeed its authors explain their rationale in the Introduction to DSLL (DSLL, Introduction: XVI-XVIII), as follows:

> Terminology is included in the approximate scope of secondary school education, in particular if [term use] is supported by evidence from journalist or popular scientific publications. The terminological entries were created partly by copying from popular science books, secondary school textbooks and specialised dictionaries, and partly by contributions from over one hundred domain specialists. Of the entire term inventory collected, only terms used in the present day were retained in the dictionary.

In the proposal for the *New Dictionary of Slovene Literary Language* (NDSLL; Gliha Komac et al. 2015: 49–51), published in March 2015, the methodological considerations concerning terminology are largely retained from the old DSLL, with some amendments. The authors of the proposal distinguish between fully and partially despecialised lexical items, whereby the former are lexicographically treated in the same manner as general words with no domain labels or counselling from experts, while the latter are to be described with simplified but scientifically correct definitions formulated by domain experts. The proposal remains vague about the distinguishing criteria between the first and second groups. The authors refer to the level of despecialisation and the familiarity of the term to general users, with both criteria to be determined from corpus data. No further details are provided about this, in our view crucial, methodological procedure.

Returning to the reasons for including terminology into general dictionaries, the despecialisation process can be frequently observed in Slovene, especially in fields such as information technology, finance, environment or sports, meaning that originally specialised terms work their way into everyday language and possibly modify their semantic and expressive scope. A general dictionary should reflect such use and define despecialised items accordingly. As for the corpus-based techniques facilitating the discovery of despecialised terms, a stratified cross-comparison of frequencies in subcorpora should provide useful clues. For example, if a term such as *infarkt* 'infarction' is found in a subcorpus of medical abstracts, but at the same time appears in general newspaper articles and user-generated contents with a different network of collocates and modifiers (*prometni infarkt, vremenski infarkt, svetovni infarkt, dolžniški infarkt*), this is a strong indicator of despecialisation and the broadening of meaning.

3 THE RATIO OF SPECIALISED VOCABULARY IN A GENERAL DICTIONARY

Assuming that users value the presence of specialised items in a general dictionary, the question that inevitably follows is how many terms should be included, or what the ideal proportion between terms and non-terms should be.

Several authors have addressed these questions, including Landau (1974), who analysed *Webster's Third New International Dictionary* and found it contained around 40 percent of terminology, with selected dictionary pages having up to 89 percent. Béjoint (1988: 360) discusses the importance of terminology, but is reluctant to specify portions or ratios because the distinction between terms and non-terms is anything but straightforward. A similar view is adopted by Boulanger and L'Homme (1991: 25), however a later study by Boulanger (1996: 147) identified between 40 and 50 percent of specialised items in English and French monolingual dictionaries. More recently, Vrbinc and Vrbinc (2013) explored the differences in the treatment of terminology in 3rd, 4th and 8th editions of *Oxford Advanced Learners Dictionary* (OALD). The authors specifically focused on the use of subject-field labels and the improvements thereof, but also found that the proportion of scientific and technical vocabulary continually increased from one edition to the next.

The approach advocated to general dictionary creation here seeks to be pragmatic in the sense that the proposed methodology for the inclusion and treatment of specialised vocabulary should be feasible in terms of time and funding, while fulfilling all the necessary criteria regarding efficiency and sustainability. An online dictionary can follow the user-centred approach in selecting which information to display to which type of user, meaning that specialised vocabulary – if properly labelled as such – need not be restrained beforehand in terms of its volume or specificity. State-of-the-art computational methods facilitate the process of extracting terms, definitions, collocations and examples from corpora, moreover they provide reliable clues about their use, frequency or variants across registers and text types. The severe space constraints which governed data presentation in the times of printed dictionaries no longer apply, and today limitations are posed not by data storage capacities or bandwidth, but by the information processing capacities of the human user.

Still, even with the best information extraction techniques and language processing tools, automatically obtained data still requires a substantial amount of validation, editing and completion by lexicographers before it can be presented to the user. Returning to the issue of the volume of specialised items to be included into general dictionaries, one should therefore note that the decision to include a

large number of terms inevitably means a large investment of human labour into this task, possibly including the involvement of domain experts.

Within the context of corpus-based lexicography, which essentially describes language use, it would therefore have been futile to predefine the amount of specialised lexical items to be included, and lexicographers might instead adhere to the relatively loose principle that the dictionary should be as specialised as necessary in order to fulfil the broadest possible range of information needs from diverse user groups, while retaining the character of a general language dictionary.

The last question we briefly touch upon is the balanced representation of domains, in other words, should a general dictionary contain equal portions of terms from all the specialised domains it includes. A review of lexicographic traditions (Josselin-Leray 2005: 146ff) shows that balance between domains was generally considered irrelevant or impossible to achieve. In many cases the reasons for a detailed representation of a particular domain in a dictionary were purely anecdotal (Béjoint 1988: 361): "One of the editors of the OED happened to be an amateur mineralogist, and consequently the [SOED] is particularly rich in words of mineralogy." Specialised domains differ in the type and number of terms they use, and some domains are certainly of more interest to the general public than others, a fact likely to be directly observable in a well-balanced reference corpus. Ahmad et al. (1995) claimed that a general dictionary should focus on the domains and terms where the layperson's and the expert's interests overlap. In the following sections we propose some methodological guidelines to help us measure this overlap and classify specialised terms accordingly, but because our approach is corpus-based we first discuss notions of representativeness and balance in the context of specialised vocabulary.

4 COMPILING A CORPUS TO EXTRACT SPECIALISED VOCABULARY: SOME CONSIDERATIONS

Since the beginnings of corpus linguistics, digital data collections have provided invaluable empirical evidence for all kinds of lexicography. For general language dictionaries the most widely used type is the reference corpus, which is often understood as a common denominator representing the broad range of language varieties and text types occurring in a language community. While the compilers of early reference corpora devoted a great deal of critical reflection to the notions of balance and representativeness, in recent years we have witnessed a trend towards compiling very large web-crawled corpora which at best represent the language of the Internet, but cannot be taken as representative nor balanced samples of the language as a whole.

For Slovene, the first reference corpus FIDA (Erjavec et al. 1998) was built much in line with the principles of the Czech or British national corpora. The currently largest reference corpus is Gigafida, containing over a billion tokens from a broad range of genres and varieties, while not claiming to be balanced. Its subset KRES, with 100 million tokens, is "artificially" balanced in that it contains equal portions of the five main text genres.

If (absolute or relative) corpus frequency is considered one of the essential criteria for the lexicographic treatment of general language, it is certainly not always reliable for specialised lexical items. While some specialised terms, for instance those pertaining to economy, finance or sports, might be well-represented in a general language corpus and thus bear witness to the above-mentioned overlap of experts' and lay people's interest, others, such as terms from the domains of math, physics or chemistry at the level of high school textbooks, will not appear as frequently.

For instance, three sibling terms (see below) occurring in a high school math textbook and designating polygons, *trapez* (trapezium), *paralelogram* (parallelogram) and *deltoid* (deltoid), are found to occur in extremely different frequency ranges in Gigafida and Kres.

	GF	Kres
paralelogram	107	18
trapez	923	126
deltoid	23	3

A general dictionary striving to include terminology up to the level of secondary education should probably contain all three, but it remains a challenge how to discern their termhood from corpus data alone. Of course, situations such as the above can easily be explained: *trapez* is a highly polysemous word which is found to have at least five other (semi)specialised meanings in Gigafida, including the domains of basketball, gymnastics, medicine, information technology and sailing. The lexicographer will very likely need to review all of these specialised meanings manually in order to decide whether they are to be included or not. In addition to polysemy, skewed frequencies may be the result of an imbalanced corpus. More specifically, Gigafida contains several years' editions of *Monitor*, a leading Slovenian IT magazine, which regularly publishes tests of IT equipment. *Korekcija trapeza* is a term frequently used in relation to screen calibration, so that the frequency of *trapez* is substantially increased due to the inclusion of *Monitor* in the data.

The above considerations bring us to the conclusion that if the dictionary has no claim to a balanced representation of various domains, but on the other hand

cannot rely solely on frequency data from a reference corpus as to which terms to include, it is necessary to identify priority domains which should be represented, and for these to provide sufficient material to allow us to exploit automatic methods of data extraction and processing. By sufficient material we mean subcorpora, of which some may already exist while others still need to be compiled.

The selection of priority domains along with the target genres and text varieties is closely related to the overall dictionary concept, and should reflect the needs and requirements of target users. However, since a thorough and extensive analysis of the Slovenian users' needs with regard to terminology has not been performed yet, these decisions will inevitably be subjective and intuitive. One attempt to overcome this issue is the analysis of Termania queries described below.

For the identification of priority domains we propose three guiding principles. The first is related to the above-mentioned intersection of expert and lay interest, meaning that the dictionary should focus on the domains which are frequently discussed in general public discourse and are well represented in the media. A quick scan through the 1,000 most frequent noun lemmata from the Gigafida corpus reveals the following list of topics: politics, sports, law, economy, finance, media, environment, administration, health, culture, IT, traffic, and tourism. It should be noted that such a scan might point us to the domains of the so-called public interest, but the lexical items occurring within the top 1,000 nouns could hardly be considered terms. Their use in despecialised contexts inevitably broadens their semantic spectrum and loosens their membership in a specialised domain.

The second principle addresses the target group of potential dictionary users in education, and at the same time aims to achieve the terminology coverage of a high school graduate. This implies that one of the essential subcorpora should consist of high school textbooks covering all subjects taught at the level of secondary education in Slovenia. Since the current version of Gigafida is imbalanced in this respect, one of the future tasks entails a systematic revision and extension of the textbook subcorpus.

The third guiding principle acknowledges the fact that even the largest and most carefully designed corpora cannot represent the entire vocabulary of a language. Leaving the huge landscape and variety of spoken discourse aside, certain areas of written communication are underrepresented in existing corpora. We have identified one such gap and labelled it broadly as life events, by which we mean a range of lexical items referring to various administrative, legal, social, religious, medical and other procedures people regularly encounter. Such vocabulary may be found in banking, insurance or administrative forms, identity cards and other types of documents associated with individual life events.

Another aspect of the third principle, which could also be referred to as the awareness of the noncomprehensiveness of corpora, is neology. New terms and expressions are coined almost exclusively in specialised domains, where a high term formation rate can be observed for the naming of new technologies, scientific findings, devices and services. Often these innovations are recieved with a wave of attention from the general public and the media. While it is difficult to follow the evolution of new terms through corpora their inclusion in the dictionary is important, especially in those cases where the newly coined term is not merely the result of a journalist's or translator's creativity, but represents the result of a term planning and harmonisation process.

One example of such a process is the quest for the Slovenian equivalent of crowdsourcing. At first the term was directly borrowed from English, and can be found in the original English spelling four times in Gigafida (crowdsourcing), then several possible translations started appearing: *moč množic, množicanje, množgančkanje, množičenje,* and *množično zunanje izvajanje*. Lively discussions about the most appropriate equivalent began and leading lexicographic institutions contributed their views, but still none of these equivalents can be found in today's Gigafida. For the new dictionary we must therefore systematically devise strategies to follow the evolution of terms and make well-founded decisions about their inclusion in the dictionary.

5 EXTRACTING SPECIALISED VOCABULARY AND OTHER LEXICOGRAPHICALLY RELEVANT DATA FROM CORPORA

Before the computer era lexicographers spent the bulk of their time building inventories of words in a language to be included in a dictionary. State-of-the-art language technologies substantially reduce this task and allow lexicographers to focus on the validation, revision and completion of automatically extracted information.

5.1 Adapting term extraction tools to the task at hand

Several tools exist for the automatic extraction of terminology from text for many languages, including Slovene (Vintar 2010). The LUIZ Term Extractor was originally been developed for bilingual extraction of terminology from English-Slovene parallel and comparable corpora, but may equally well be used for Slovene

alone. The underlying assumption of many term extraction methods is the notion of keyness (Scott 1997), which compares the frequency of a selected lexical item in a specialised corpus with its frequency in a general language corpus, and assumes that if the item is relevant or "key" to the specialised domain it will occur in the specialised corpus with a higher relative frequency than in the reference corpus. In LUIZ, the "keyness" of a term candidate is combined with part of speech patterns and a ranking heuristic to provide a list of candidate single- and multiword terms as output.

LUIZ has been tested for various domains (Vintar and Erjavec 2008; Vintar in Fišer 2009; Vintar 2010; Logar et al. 2012; Pollak 2014), but has never been used as a tool to extract terms for a general language dictionary. Several extraction parameters need to be adjusted to the task at hand, such as:

- The lenght of extracted terms. While for specialised terminography target units may contain three, four or more words, for a general dictionary it is better to focus on less specialised, therefore shorter terms containing one or two, and exceptionally three, words.
- Part-of-speech patterns. For specialised dictionaries the typical morphosyntactic term patterns may vary, but usually we attempt to achieve maximum recall by specifying all potentially productive patterns for a given language. Here, the great majority of specialised items is expected to be either single nouns or two-word combinations of adjective+noun or noun+noun, because in previous experiments these two patterns have proved to be most productive in term formation.
- Ranking heuristics. In a specialised terminography task it is not uncommon to extract units occurring only a few times, while for a general dictionary we tend to avoid items which are too specialised or rare.
- The definition of subcorpora for the computation of keyness. Keyness works if a clearly delimited specialised corpus is compared to a much larger reference corpus. In the context of our dictionary project, subcorpora will need to be defined for each individual extraction task. The textbook subcorpus is for instance entirely unsuitable for term extraction, because it contains a number of domains which will level each other out.

The result of multiple rounds of term extraction and cross-comparisons between subcorpora is lists of candidate terms for selected domains requiring thorough validation and supplementation. This will involve the identification of new terms which may have been overlooked by the term extractor due to low frequency, but also the ordering of terms into concept networks which will help identify gaps, near-synonyms and missing hyper- and hyponyms.

This stage of specialised vocabulary compilation already requires the involvement of domain experts to help identify term variants, obsolete or non-standard terms and synonyms, which will facilitate the lexicographers' decisions regarding classification into various groups of lexicographic treatment.

5.2 PILOT EXPERIMENT: EXTRACTING TERMS FROM TWO SUBCORPORA

In order to get a clearer insight into issues related to the representativeness of subcorpora and the necessary adjustments of the term extraction tool, we performed a pilot experiment. The LUIZ term extraction tool was used on two subcorpora, one containing a selection of texts pertaining to physics, biology and chemistry from the ccGigafida corpus (Logar Berginc et al. 2012: 77–97), and the other a more homogeneous specialised corpus of textbooks on music theory.

The subcorpus of natural science (Nature) is composed entirely of texts already included in ccGigafida and contains 13 primary or secondary school textbooks on natural science, biology, physics or chemistry, and 16 other popular scientific books from various publishers on the topics of astronomy, botany and gardening. We also included texts from related magazines including educational and popular-scientific periodic publications (*Gea* [geography], *National Geographic* [geography], *Kmetovalec* [agriculture], *Moj lepi vrt* [gardening], *Mrgolazen* [biology], *Ribič* [fishing], etc.).

The music subcorpus (Music) contains 10 contemporary textbooks used for musical education at the level of primary and secondary education. The corpus was compiled as part of a PhD study by Jelena Grazio at the Department of Musicology at the University of Ljubljana, and the textbooks deal with diverse areas of music theory (harmony, solfeggio, counterpoint, music forms). It is important to note that none of these textbooks had been part of the Gigafida or ccGigafida. Table 1 lists basic information about the two subcorpora.

Table 1: Basic information about the Music and Nature subcorpora

	Music	Nature
Tokens	280,060	1,053,897
Types	12,121	59,788
Number of documents	10	388
Text variety	textbooks	textbooks, popular-scientific books, magazines

For the term extraction experiment we used LUIZ with modified settings, more specifically we limited the extraction patterns to single nouns and adjective+noun phrases. Termhood is computed using the LUIZ heuristics, and is based on measuring keyness against the entire Gigafida corpus. Table 2 demonstrates the differences between subcorpora in size and the number of candidates extracted, with 9.3% single-word term candidates in Music and 13% in Nature, while two-word candidates seem to be more common in Music (6.8%) than Nature (2.2%).

Table 2: Number of term candidates extracted from both subcorpora

	Music	Nature
Noun	1,137	7,853
Adjective + Noun	828	1,309

The higher percentage of two-word terms in Music already highlights one important difference between the subcorpora – the level of specialisation, and another difference becomes apparent when we inspect lists of term candidates and observe a high level of domain homogeneity in Music, while the Nature corpus is much more diverse with regard to domains and topics. We analysed the top 150 term candidates from both single- and two-word lists and from both subcorpora and arrived at the following conclusions:

- The lists of terms from Nature reveal an imbalance between different text domains and sources, so that terms from certain domains seem overrepresented (such as fishing and gardening). For the purposes of term extraction subcorpora should ideally be as homogeneous and balanced as possible.
- The term candidates from the Music subcorpus contain a relatively low proportion (about 30% single-word nouns and 15% two-word candidates) of entirely despecialised terms which can be considered part of general vocabulary and require no special lexicographic treatment (e.g. *takt, glas, nota, melodija, skladba, harmonija etc.*; *notno črtovje, klasična glasba,* and *klavirska spremljava*). All the other candidates are specialised terms clearly belonging to the musical domain and not necessarily understood by lay persons (e.g. *modulacija*, *fuga*, *kvintakord*; *tritonusna kvinta*, *eolska septima*, and *napolitanski sekstakord).*
- With the Nature subcorpus, the situation is reversed: a large majority of single-word nouns refer to general concepts with a very vague affiliation to a specialised domain (*rastlina*, *voda*, *list*, *seme*, *plod*, *poganjek*, *temperatura*, *svetloba*; *okrasna rastlina*, *soška postrv*, *organski odpadek*, and *listna uš),* and the list of two-word candidates contains only about 15% of terms which might require a more technical definition (e.g. *ogljikov*

hidrat, *celična membrana*, *magnetno polje*, *maščobna kislina*, *potencialna energija*, and *vrtilni moment).*

It would appear that a more specialised and homogeneous corpus is more appropriate for the term extraction task, but on the other hand a general language dictionary need not contain highly specialised terms unless they fulfil the inclusion criteria discussed above. Textbooks and magazines certainly represent valuable sources of terminology, but their lexicographic description will depend on their degree of specificity which we discuss in more detail in the following section.

6 DEGREES OF SPECIFICITY AND LEXICOGRAPHIC DESCRIPTION

As illustrated by the examples above, terms vary in their degree of specificity. The goal of the new dictionary is to satisfy at least three diverse groups of users: learners at all levels of education, language professionals and those having language – including its peculiarities and specialised expressions – as a hobby, and as a consequence lexicographic descriptions should be tailored both to the needs of potential dictionary users and to the properties of the lexical item itself.

In line with these considerations we propose the categorisation of specialised items into three groups or baskets, each requiring a different approach to lexicographic description.

The first is the general basket, which contains the least specialised items. As such these items may exhibit a vague relation to a specific specialised domain, however knowledge of the domain is not a prerequisite for understanding or use. Such lexemes typically occur in the reference corpus with a relatively high frequency (> 3,000), and are used in general texts entirely devoid of domain-specific references. The lexicographic description may be generic, without domain labels or technical definitions, and theinput of experts is not required. Examples of such items from our Music subcorpus might include *tempo, koncert, dirigent, nota,* and *harmonija*.

The second is the so-called school basket, and contains terms less frequently encountered in general texts, although they still designate basic concepts of the domain and may already occur in textbooks at the level of primary education. Their membership in the specialised domain is clearly identifiable, however they do occur in the reference corpus (> 300). A lexicographic description may contain a domain label within the gloss, especially in cases where the despecialised meaning deviates from the domain-specific one. Examples include *sonata, akord, dur, mol, oboa,* and *trozvok*.

The third is the so-called technical basket, containing truly specialised lexical items less familiar to the general public and requiring some knowledge of the domain. They seldom occur in general texts and are not used with a despecialised meaning, although they may be considered for inclusion into the general dictionary for various reasons: either they occur in secondary school textbooks, and thus represent the vocabulary of an average high school graduate, or they have been found to belong to one of the priority domains and occur in the reference corpus with a minimum frequency. The lexicographic description will reflect the degree of specificity and should contain a domain label, the definition should be a (possibly simplified) terminological one, formulated or validated by a domain expert. Examples of such terms from the musical domain include *septima, alikvotni ton,* and *sekstakord*.

If the lexical item does not fulfil any of the above criteria for inclusion, meaning that it does not pass the threshold frequency in the reference corpus, is not part of a priority domain and does not occur in a textbook, nor does it represent an indispensable part of the conceptual network of another – more frequent – term, we may exclude it from further treatment and assume its degree of specificity to be too high for a general language dictionary.

The most challenging part of the proposed classification is the efficient and reliable exploitation of corpus data, since – as illustrated above – sheer corpus frequency cannot be seen as a reliable indicator of termhood. A common reason for skewed frequencies is ambiguity, where lexical items may have specialised meanings in several domains. An example is the musical expression *sinkopa*, which occurs 269 times in the Gigafida corpus, although the majority of occurrences pertain to the medical meaning of the term (a temporary loss of conscience). A number of occurrences include *sinkopa* in names of companies, musical groups and products, and only a small number represent the meaning in the musical domain. Since reliable word sense disambiguation tools still have not been developed for Slovene, ambiguity represents the largest obstacle to exploiting raw corpus data. This problem is largely resolved by using domain-specific subcorpora, and by cross-comparisons between them we may arrive at more realistic conclusions about the frequencies of individual meanings.

7 ANALYSING TERMANIA SEARCH QUERY LOGS TO ASSESS POTENTIAL USERS' TERMINOLOGY NEEDS

The new dictionary is ambitious in that it attempts to satisfy a broad range of target users whose communicative actions are increasingly embedded in the digital world. While a number of studies deal with user scenarios and their expectations

for online dictionaries, few studies focus specifically on users' needs with regard to terminology. An important contribution has been made by Amelie Josselin's PhD thesis (Josselin-Leray 2005), which investigates the role of terminology in general dictionaries from various aspects, including conducting surveys to identify what dictionary users want and need as far as terms are concerned. The empirical results of these surveys are summarized in Josselin-Leray and Roberts (2007), and among other findings reveal that users place a high value on the exhaustiveness of a dictionary, that they generally expect monolingual dictionaries to contain more terminology that bilingual ones, but also that French users on average place a significantly higher value on the presence of terminology in dictionaries than English users.

To date, no similar study has been performed in Slovenia, but in order to shed some light on user information needs regarding specialised vocabulary we performed an analysis of search query logs for the Termania.net dictionary portal. Termania.net is a free dictionary aggregation portal created by Amebis d.o.o. in 2010, and represents one of the largest online resources for terminology from various domains. The portal provides unified access to 44 dictionaries, of which approximately half are categorised as general. These include access to the DSLL, the Slovene lexical database, various dictionaries of rhymes, abbreviations, dialects and several bilingual dictionaries. Specialised resources include small, medium and large dictionaries from numerous domains, including education, medicine, biology, IT, and tourism.

Since the Termania.net site is very well-known among translators, students and the general public, its search query logs might provide useful insights into the information needs of Slovene users. For the purpose of this analysis, Amebis d.o.o. provided logs for the past two and a half years ordered by frequency. The number of all queries was over 6.5 million, with 433,692 different ones, of which 287,283 occur only once. Unfortunately the logs do not tell us whether the query was successful nor do they reveal any information about the users.

Our main interest was to see whether Termania users search for terminology, but also to inspect query types (single- or multi-word, use of wildcards, etc.) and gain insights into their information needs. Since it would have been impossible to manually inspect all of these we automatically compared the list to the following resources:

- the Gigafida reference corpus of Slovene,
- the EN1010 web-crawled corpus of English,
- the first edition of the *Dictionary of Slovene Literary Language* (DSLL1),
- the second edition of the *Dictionary of Slovene Literary Language* (DSLL2).

Table 3: Checking the presence of Termania search queries in other resources

	Število poizvedb	Odstotek poizvedb
Termania	**433.692**	**100%**
Termania - 1x	287.283	66.3%
is in GF	177.687	40.1%
is in DSLL1	61.393	14%
is in DSLL2	64.348	15%
is in EN1010	114.044	26%

Table 3 presents the results of these comparisons. Our first observation was that over a quarter of all queries can be found in the English corpus, which leads to believe that Termania is frequently referred to as a bilingual resource. About 40 percent of queries can be found in Gigafida, but since the comparison was performed with the list of lemmata, this number might be slightly higher because users occasionally search for inflected word forms. Only 14 percent of different queries can be found in the DSLL, although the sum of these queries amounts to 2.6 million, which is around 40% of all queries. As many as 94,891 queries are multi-word, of which the majority are terms.

The queries not occurring in any of the compared resources can be categorised as follows:

- Slovene words in inflected forms, rarely terminological: *podatki, iščem, priljubljena, spremljaj,* and *zadržano*
- English words in inflected forms: *prospecting, sieving, levying, inventorying,* and *garnishing*
- Slovene specialised terms, especially borrowings: *hipersoničen, ekhimoza, acianotičen, transudat, distenzija, mezotelij,* and *hipersplenizem*
- Queries containing an asterisk: *an*, k*, turist*, pos*, spoln*,* and *hidroksi**
- Searching for word suffixes using a hyphen (not supported by the Termania search engine and therefore always unsuccessful): *-olg, -okate, -njiva,* and *-njice*
- Abbreviations, acronyms: *crh, ToR, ZAZV, Tfc,* and *accn*
- Words from languages other than Slovene or English: *einrichtung, belegen, verteilen, stellung, ausgleich, Spannungsversorgung, knikken, gebruiken, plutajuči,* and *το θηρζον*
- Other: proper names, expressions in brackets or quotation marks, misspellings, numbers, symbols, and nonsensical character strings.

The analysis shows that terminology constitutes an important part of all queries on Termania, and thus that users frequently refer to an online portal to resolve their information needs. Cross-comparisons between different resources provide important clues for the new dictionary, for example by exploring queries not present in any of the Slovene dictionaries, but coming up in Gigafida with several hundred occurrences.

8 CONCLUSIONS

The purpose of this chapter was to explore the role of terminology in the new *Dictionary of Modern Slovene Language* from as many aspects as possible, first by positioning it with regard to existing lexicographic traditions on the one hand, and the aims of the new dictionary on the other, then by setting up a methodological framework of corpus-based terminology acquisition for the purposes of a general language dictionary, and finally by reflecting on the categorisation and lexicographic treatment of terms, in light of the target users' needs.

Clearly, to create an exhaustive dictionary with a high coverage of specialised terminology is an ambitious goal in itself, and to achieve it the methodological considerations presented above will need to be continually refined, improved or modified, at all stages of the process. A concern for Slovene specialised language is a recurring theme in the national language policy, and has found its way into Slovene legislation and higher education. We hope that the *Dictionary of Modern Slovene Language*, with its comprehensive and non-exclusive coverage of terminology, will primarily help users communicate knowledge.