

Building lexica for speech recognition

Agnieszka Wagner

Department of Phonetics
Institute of Linguistics, Adam Mickiewicz University
wagner@amu.edu.pl

1. General information

The main goal of developing the lexica presented in this report was creation of language resources suited for application in speech technology systems and especially in a speech dictation and recognition system *Jurisdic*.

For this reason, part of the language resources described in this report has been created according to the LC-STAR project specifications [1]. In the project three lexica have been proposed covering: *common words*, *proper names* and *special application words*. The common words (CW) lexicon should consist of at least 50,000 entries extracted from large text corpora in order to cover a broad range of domains. The proper names (PN) lexicon should include at least 45,000 proper names extracted from various domains relevant for speech recognition and speech synthesis (this lexicon is presented in a separate report). The special application words (SAP) lexicon should include at least 5,000 special application words relevant for voice driven applications.

Apart from the lexica proposed in the LC-Star project an additional lexicon based on frequency dictionaries has been created in order to ensure high lexical coverage: the procedures used to built it are described in detail in the report.

2. Description of domains and sources for the common word (CW) lexicon

For the creation of the lexicon a text corpus had to be built based on sources/media described in sec. 2.2 and presented in table 1. The corpus includes 177.64634 tokens ('cleaned' data, see table 2) in order to ensure high lexical coverage of a broad range of domains. The covered domains are described in the following subsection (2.1). From these domains a word list was extracted after text normalisation (for detailed information see: [2]). The criteria for statistical entry selection is given in sec.3.2. The LC-Star requirements specified that the final lexicon should consists of at least 50,000 entries extracted from the corpora in order to cover a broad range of domains.

The type of linguistic information that the lexicon and other lexica has been enriched with is given in sec. 6.

1. Lexical domains

The table below presents application domains and exemplary text sources/media chosen for the collection of the common word corpus.

Table 1: Application domains for the common word (CW) lexicon

Domains	Subdomains	Source/media
C1. Sports/Games	C1.1.Sports (special events)	Medycyna sportowa Rzeczpospolita, (section: sport)
C2. News	C2.1. Local and international affairs	Rzeczpospolita, Wprost (section: country, the world),
	C2.2. Editorials and opinions	Rzeczpospolita (section: opinions)

C3. Finance	C3.1. Business, domestic and foreign market	Rzeczpospolita (section: economy), Wprost (section: business)
C4. Culture/Entertainment	C4.1. Music, theater, exhibitions, review articles on literature	Esensja - www.esensja.pl/ Rzeczpospolita (section: culture), Teatry - www.teatry.art.pl/
	C4.2. Travel / tourism	Rzeczpospolita (section: travels, holidays), Magazyn Świat - www.magazynswiat.pl
C5. Consumer Information	C5.1. Health	Medycyna sportowa, Gazeta Lekarska, Nasze Miasto (section: health)
	C5.2. Popular science	Rzeczpospolita (section: technology, car)
	C5.3. Consumer technology	manuals, instructions for desvices and appliances from Panasonic, Nokia, Samsung, Sony etc.
C6. Personal communications	C6.1. Emails, online discussions, editorials, e-zines	online discussion fora of the newspapers: Życie Warszawy, Wprost, letters to the editor

2. Sources/media

For corpus creation electronically available text corpora were chosen: newspapers, periodicals, manuals, data from the internet (online magazines, newsletters, discussion groups, newsgroups). To assure that all domains are represented, each domain is represented by at least 1 Mio of tokens (i.e. after text cleaning; cf. tokenization procedure, see statistics in table 3). The requirements considering the cut-off date for all corpora used (i.e., the year 1990) and the year of publication (i.e., at least 50% of the newspapers /periodicals used should have appeared after 1997) have been met. All the text sources have been used with the agreement of the publisher/author. Apart from the abovementioned requirement, the selection of text sources has been based also on another criterion, namely the format of the text. Since determination of the lexical domain the specific text represents could not be carried out manually due to the large amount of data, it has to be based on information available directly from the text. For that reason only those resources which offered text in the HTML-code were taken into account in the corpus creation. So, the decision whether a specific text sample represents a particular lexical domain or not was based solely on the HTML-tags; the example below comes from the online edition of the daily Polish newspaper Rzeczpospolita:

```
href="/gazeta/wydanie_XXXXXX/publicystyka/  
href="/gazeta/wydanie_XXXXXX/kraj/  
href="/gazeta/wydanie_XXXXXX/swiat/  
href="/gazeta/wydanie_XXXXXX/kultura/  
href="/gazeta/wydanie_XXXXXX/nauka/  
href="/gazeta/wydanie_XXXXXX/sport/  
href="/gazeta/wydanie_XXXXXX/ekonomia/
```

This approach was especially useful in case of daily newspapers (such as the abovementioned Rzeczpospolita or Głos Wielkopolski) and magazines (e.g., Polityka, Przekrój) which offer a number of sections dedicated to different topics including economy, sports, culture, etc. As regards specialist magazines (e.g., Przegląd Sportowy - Sports' Digest or Gazeta Lekarska - Medical Newspaper) it was assumed that all the text material they included was representative for the lexical domain determined by their profile (thus sports and health for the two exemplary magazines just mentioned).

As specified in the LC-Star project in order to get a maximum of lexical entries an overlap of entries within the common domains, the proper names domain and the special application domain should be avoided as much as possible: the final lexicon has to consist of at least 100.000 different entry groups. For that reason, after a pre-processing of texts (see [2] for details) a preliminary morphological analysis was carried out [3] in order to identify paragraphs including a large number of proper names which were later eliminated from the corpus. The cleaning of the CW corpus of the remaining proper names and abbreviations was carried out during tokenization. The CW list was also verified manually: we searched for proper names and abbreviations, deleted them from the list manually and added to the PN and SAP lists respectively (see sec. 3.2). All entries included in the SAP list which were found also on the CW list were removed by hand and replaced with new entries.

3. From text corpora to word list

In the following sections the strategies for creation and extraction of the common words list are described.

1. Word list creation for the CW lexicon

As specified in [1] in order to get a good lexical coverage for the common word domains the word list for each domain must reach a target of at least 95% self coverage on the common words of the corpora used for this domain and the final word list of all common word domains together must reach a target of 95% self coverage on the common words of the whole corpus. In addition to the coverage criteria at least 50.000 different entries should be provided.

The assignment of text material to specific domains was based on the information available in the HTML-code as described in sec. 2.2.

2. Word list extraction for the CW lexicon

The procedure for word list extraction is described below.

The list of common words is created from the collected text corpora, so that:

1. The most frequent words are present. This requirement has been checked by comparing the overlap between the final CW list and frequency list (including all inflectional forms generated for the 23624 most frequent Polish words, see sec. 5). The comparison showed that the overlap reaches 85% (i.e., this percentage of the words present on the CW list can also be found on the frequency list), which confirms that the CW list includes the most frequent words.
2. Each of the six domains is covered to at least by 95% (disregarding singletons). This requirement has been met (see p. 4 below).
3. The total number of common words is at least 50,000. This requirement has been met: the final CW list includes 92607 entries (see p. 12 below).

Formal procedure for word list extraction:

1. The collected corpora have been cleaned and tokenized for all domains: digits, punctuation marks, typos were removed. The tokenization procedure is given in [2].
2. The size requirements on domains have been verified: it was assumed that a good lexical coverage can be achieved when each domain is represented by at least 1 Mio of tokens. In our corpus each domain is represented by more than 1 Mio of tokens (see table 2 below).
3. Proper names and abbreviations have been removed as much as possible by automatic means. Preliminary analysis of the text source data have shown that some types of texts are rich in proper names, e.g. in the sport magazines whole paragraphs consisted of list including names and results (digits and numbers). Therefore, after a pre-processing of texts (see [2] for details) a preliminary morphological analysis was carried out in order to identify such paragraphs and eliminate them from the CW corpus. The cleaning of the CW corpus of the remaining proper names and abbreviations was carried out during tokenization.
4. As proposed in [1] the basic coverage target $t = 95\%$ has been increased: we used $t = 96\%$. A preliminary analysis of the word lists have shown that there are still some proper names and abbreviations left on the lists which had to be removed manually. For this reason, in order to meet the coverage requirement a safety margin of 96% has been applied.

The table below shows general statistics of the corpora and word lists:

Table 2: Overview of text corpora and word lists for the CW lexicon

DOMAINS	CLEAN CORPORA		DIFFERENT_WORDS	
	tokens	distinct tokens	number	coverage
Sports/Games	2320749	54609	23099	96%
News	3828706	85412	40838	96%
Finance	1680426	40031	17602	96%
Culture/Entertainment	4309706	110665	49530	96%
Consumer Info	5625047	111311	43502	96%
Total	177.64634	402.028	78.150	

The word frequency in each domain has been determined as follows:

5. The number of occurrences of all distinct tokens (words) in the corpus has been count and the words have been sorted with the decreasing frequency.

6. Singletons have been removed
7. Relative frequencies $Rf(w)$ have been calculated: the frequencies of all words have been summed up $Sf(w)$ and then each observed word frequency $Of(w)$ has been multiplied by 100 and divided by $Sf(w)$
8. self-covarege (SC) of the word list has been determined by summing up $Rf(w)$:

$$\text{if } Rf(w_i) > Rf(w_j), \text{ then } SC(w_j) = Rf(w_j) \setminus 100 * Sf(w) + Rf(w_i) \quad (1)$$

9. each CW list was truncated at the point where $SC(w_j) = 96\%$
10. the lists have been verified: proper names and abbreviations have been identified and removed manually as much as possible
11. the steps 11 and 12 have been iterated
12. the word lists for all domains have been marged: the resulting CW list includes 92979 entries

The final CW list required manual verification: remaining proper names had to be identified and hand-deleted. Apart from few cases of words which undoubtedly should be deleted from the final CW list (e.g. Bhutan, Damaszek, Gardno, Giewont) and resulted from errors in the preprocessing and tokenization of the text corpora, the list included a number of words that 1) could function as both proper names or common words: the distinction depends only on the capitalization of the initial letter e.g., *kandahar* (type of a gun) vs. *Kandahar*, or *kozłowski* (an adjective derived from the town name Kozłów) vs. *Kozłowski*. The list was searched for this kind of words and they were marked. We have also marked 2) foreing names (to pay special attention to them while checking the results of automatic annotation), 3) trademark names (e.g. Opel, Fiat - since they can be written with a capitalized initial letter or not depending on how they are treated - to make sure that they also appear in the PN lexicon) and 3) words which often constitute part of proper names e.g., *zdrój* (literary term for spring) vs. *Jastrzębie Zdrój* (name of a popular Polish spa town). The decision had to be made whether words in 1) and 4) should be left or rather removed from the CW list. Since the distinction between proper names vs. common words is made on the basis of capitalization it has to be decided whether a given word should rather be treated as a PN or CW. This problem could be left to be solved in the linguistic analysis module of the recognition system, but since this analysis has some restrictions and is not always 100% successful we preferred to avoid future errors. Initially, 759 words have been selected and it has been checked with a number of dictionaries whether they function solely as PNs or also as adjectives derived from other PNs (Chabowski vs. chabowski -> Chabówka, Jabłoński vs. jabłoński -> Jabłonów, Zakrzewski vs. zakrzewski -> Zakrzewo). Words for which no corresponding entry could be found in the dictionaries have been moved to the proper name list. If there were any doubts, we have searched for the word frequency in google. It has been observed in which source a given word has been found and in which context. On the basis of this information words have been either deleted from the CW list or left. The domains in which the words frequently occurred in the corpus also provided a very important information. Here are some examples of the deleted words:

- łokietek: diminutive of 'łokieć' (elbow), occurs rarely in this meaning but at the same time is the name of Polish king -> *Władysław Łokietek*. Additionally, the word was present only in the *culture* domain.
- nagano: vocative of 'nagana' (reprimand); vocative is very rarely used for nouns other then those belonging to semantic category *person* (teacher, accountant, friend, etc.) but *Nagano* is the name of a very well-known city, a ski-jumping centre. Additionally, nagano has ocurred only in the *sport* domain.
- rydzyk: diminutive of 'rydz', (saffron milk cap): again very unusual in this form but at the same time - the surname of the head of Congregation of Redemptorist Fathers in Poland, Tadeusz Rydzyk - a very well known and controversial person, very often present in the media. Additionally, rydzyk appears only in the *news* domain.
- rokita/rokite (inflected 'rokita'): a plant species (type of a willow), when capitalized: the surname of a Polish politician Jan Maria Rokita. The word occurred only in the news domain which clearly indicates that it should be deleted from CW and moved to the PN lexicon.
- poznań: plural, gentivive of 'poznanie' (cognition). This word occurs rarely in this form but at the same time it is the name of one of Polish major cities: Poznań.

Altogether, 373 words of this type have been deleted from the CW list and if not present - added to the PN list in the lemmatized form. The final CW list includes 92607 entries. The top of the list is given in the table below. Like in other languages (see [1]) it includes most of all function words: pronouns (to, się), prepositions (za, na, do, po), conjunctions (ale), particles (nie) and auxiliary verbs (jest, są).

Table 3: Top of the frequency list for the common word lexicon

word	total frequency	culture	consumer information	finances	sport	news
na	358378	90988	111703	35996	51200	68491
się	337678	89580	110699	25978	47103	64318
nie	263141	56361	84588	19971	38468	63753
do	231612	52070	74399	23385	29497	52261
ze	169913	32780	49642	17197	22513	47781
to	156662	37699	47639	12726	21171	37427
jest	150408	37057	55756	12965	13022	31608
po	79518	20125	23454	5674	15546	14719
od	78515	19788	25578	9536	8445	15168
są	63061	16205	24536	5848	4847	11625
za	62167	15405	17707	7026	7623	14406
ale	61931	16094	17141	4101	12170	12425

4. Description of domains and sources for the special application domain (SAP) lexicon

The special application domain lexicon contains 5,177 entries (see table 5) and is further subcategorized into semantic domains described below. It consists of two parts: numbers, letters, abbreviations extracted from the CW corpora and a specific vocabulary for applications controlled by voice (information retrieval, controlling of consumer devices, etc.). The specific vocabulary list has been created by hand by a linguist on the basis of various electronically available sources described in the sec. 3.2.

1. Lexical domains

The table below presents application domains and exemplary word list entries of the special application domain list.

Table 4: Application domains for the special application words (SAP) lexicon

Common domains to be useful for all applications	Subdomains	Example
0.1. Numbers and digits	0.1.1. Cardinals (all components which can not be derived by rules should be covered)	zero to nine, eleven, twelve ... nineteen, units of tens (twenty... ninety) hundreds, thousands millions (1-31)

		first, second, twentieth thirty firstfifty-second a, b, c, ...z
0.2. Abbreviations	0.1.2. Ordinals (only those used with dates and weeks of year) 0.1.3. Letters Abbreviations from common word corpus	
1. Global domains		
1.1. Measures	1.1.1. Measures of length 1.1.2. Measures of weight 1.1.3. Measures of time and date 1.1.4. Measures of capacity	mile(s), meter, kilometer, inch, foot, pound, kilogram, hour, year
1.2. Abbreviations	1.2.1. Web-based abbreviations/parts of sites 1.2.2. Document type abbreviations	.com, .org, .gov, .edu, .pdf, .rtf, .doc
1.3. Special signs	1.3.1. Special (keyboard) signs	@, <, >, *, #, +, -, (,), :, ;
1.4. Domestic equipment	1.4.1. Household	refrigerator, table
1.5. Health	1.5.1. Common parts of body 1.5.2. Common illness	body, head, face cold, infection, cough
1.6. Greetings	Official and colloquial forms	good_morning hi
2. (Information) Services/ Retrieval		
2.1. Financial/Commerce	2.1.1. (Home Banking-(transactions) 2.1.2. Credit cards 2.1.3. Currency (list of currencies from the internet) 2.1.4. Stocks	-) withdraw, savings, balance, fee interest, credit, Euros, Dollars, Yen, local currency, cents shares, stocks
2.2. Billing	2.2.1. Payment, prepayment, payment by other means 2.2.2. Order forms 2.2.3. Sending information by other means	owe, balance, fee, unit_price, check check, required, optional, order e.g. (snail)mail, fedex, express
3. Travelling		
3.1. Tourist	3.1.1. Public (mass) transport 3.1.2. Attractions 3.1.3. Accommodations 3.1.4. Restaurant (bar) types (types, countries, sub-types) 3.1.5. Food	connections, train museum, gallery, hotel, first_class, second_class, motel, restaurant, bar, café, Chinese, Italian, vegetarian, fast_food, kosher etc. beverages: basic food
4. Information Services		
4.1. Services (general scenarios)	4.1.1. Ordering tickets	tickets, showing,

	4.1.2. Purchasing items	sizes, colors, etc.
	4.1.3. Reservations	plane, theater, movie
	4.1.4. Job Placement	shifts, night
	4.1.5. Form Submission	Enter, search, submit
	4.1.6. Web-based services	search_engines, (music) server
5. Retrieval/Control		
5.1. Retrieval	5.1.1. Horoscope	Aquarius, Capricorn
	5.1.2. Weather (weather forecast, sea weather)	temperature, highs
	5.1.3. Sports	basketball, football
	5.1.4. Traffic	highway, freeway,
	5.1.5. Personal Identification	height, weight,
5.2 Control	5.2.1 Car navigation	directions, window, door, open, close
6. Telecommunications		
6.1. Telecom	6.1.1. E-mail	send, cc, reply (all)
	6.1.2. Voice-mail	store, greeting, play,
	6.1.3. Call control	pre-paid, forward
	6.1.4. Business assistant	call, get, remind
	6.1.5. SMS	send, write, get
6.2. Web/Internet	6.2.1. Navigation in webpage	edit, delete, save,
	6.2.2. Peripherals	RAM, port, modem,
	6.2.3. Homepage	link, home, open,
	6.2.4. Voice_portal	FAQ, homepage,
	6.2.5. Browsers/servers	search, find, connect
	6.2.6 Customer Support (restricted to telecommunication)	bugs, items, broken, software, memory

2. Sources/media

Text sources/media chosen for the creation of the SAP lexicon were not specified in detail in the LC-Star project. In some cases (as regards various SAP subdomains) there was no need to refer to any specific source (e.g., 0.1. Numbers and digits or 5.1.1. Horoscope): it was enough to rely on the linguistic competence of the person in charge of vocabulary collection. Yet, whenever necessary various sources/media have been referred to including thematic dictionaries, technical documents and web portals. Here are some examples:

- The great dictionary of abbreviations and acronyms [4] (domain: Abbreviations 0.2, Measures 1.1)
- <http://www.acronymfinder.com>, <http://www.abbreviations.com> (domain: Abbreviations 1.2)
- <http://www.medycyna.szukanie.info/>, <http://www.info-med.pl/katalog/leksykon/terminologia/>, A concise medical dictionary [6] English in medical practice [7] (domain: Health 1.5)
- <http://www.bankier.pl/slownik>, <http://www.kalkulatorpodatkowy.pl/slownik>, <http://www.finance.egospodarka.pl>, <http://www.ekonom.info> (domain: Financial/Commerce 2.1, Billing 2.2)
- <http://foldoc.org/contents.html>, www.wikipedia.pl, Dictionary of IT terms [9] (domain: Web/Internet 6.2)

- The dictionary of 100.000 necessary words [5], Universal thematic dictionary [8] (domains: Domestic equipment 1.4, Travelling 3, Services - general scenarios 4.1)

Table5: Distribution of vocabulary in the SAP lexicon domains.

DOMAINS	NO. OF ENTRIES	PERCENTAGE
Numbers/Digits	167	3,23
Abbreviations	601	11,61
Global domains	894	17,27
(Info.) Services/retrieval	436	8,42
Travelling	1070	20,67
Info. services	597	11,53
Retrieval/control	975	18,83
Web/Internet	437	8,44
Total	5177	100

5. Description of procedures for creation of the frequency lexicon

Since Polish is a highly inflectional language an additional lexicon has been built in order to ensure high lexical coverage of the developed language resources.

1. The dictionary list

The first step in the building of the lexicon involved creation of a frequency list based on existing (frequency) dictionaries of Polish [10],[11],[12]. Altogether 30572 lexemes were collected. The list was examined by three linguists (with experience in lexicography) in order to eliminate less frequent words (e.g., old-fashioned words or occurring only in narrow lexical domains). During selection of the words, a frequency list created for *TranslatICA* machine translation system and based on the biggest available Polish text corpus [13] was used as a reference. It was checked which words overlap and those which had the lowest frequency on the *TranslatICA* list have been deleted from the dictionary list. We also made sure that all the words present at the top of the *TranslatICA* list occur on the dictionary-based list too and missing words have been added. The resulting list consisted of 25.757 lexemes.

Since the list included lemma, the next step consisted in generating all grammatical forms for the lemma, which resulted in 570.243 new items. As the final goal is to determine the frequency of the words on the inflected dictionary list, the following steps have been taken:

1. Verification of the common word list. We have checked the overlap between the CW and inflected dictionary list in order to ensure that the former includes the most frequent words as required in the LC-Star project. We found that 85% of common words is present also on the frequency list which proves that the LC-Star requirement is met.
2. Removal of the overlapping words from the dictionary list: since they occur in the CW list their frequency is already known.
3. Pre-selection of less frequent words. A preliminary analysis of the inflected dictionary list showed that words may occur frequently in specific grammatical forms but at the same time rarely in others. For example, as regards nouns in the masculine, a distinction is made between masculine personal and inanimate, but no such distinction is made for nouns in the feminine. It would be useful to have such a distinction, because inanimate nouns in the feminine are very unusual in the vocative case (e.g. *pasjo* - passion, *pasieko* - apiary). On the other hand, sometimes it is the grammatical form itself that is very rare irrespective of the word semantics. Almost all verbs in the imperative ending with *-ze* (*pchnijże* - let's push), *-yż* (*paplajmyż* - let's babble) or *-eż* (*pamiętajcież* - let's remember) are rarely in use. So, on the basis of grammatical information words similar to those presented here (with respect to frequency) have been marked (the final decision was made by two linguists) and filtered out: they constituted ca. 10% of the entries on the inflected dictionary list.
4. The frequency for the remaining words have been searched in Google by means of a Python script. Initially, it was intended to first check with Google the frequency for a sample of 1000 of the pre-selected words in order to ensure that they indeed represent less frequent words. Yet, information provided by the absolute frequency would be not very useful, because we still do not know how the frequency is distributed: it could be seen that in the common word

lexicon some entries had relatively low frequency, but still were present on the list. Therefore, we decided to rely on the competence of the linguists and to get back to investigate this issue later.

- In order to shorten the time necessary for the word frequency search, once 50% of the missing frequencies were found a preliminary analysis of the frequency distribution was made. We found out that it is very unequal like in the case of the CW list. This is illustrated in the figure below:

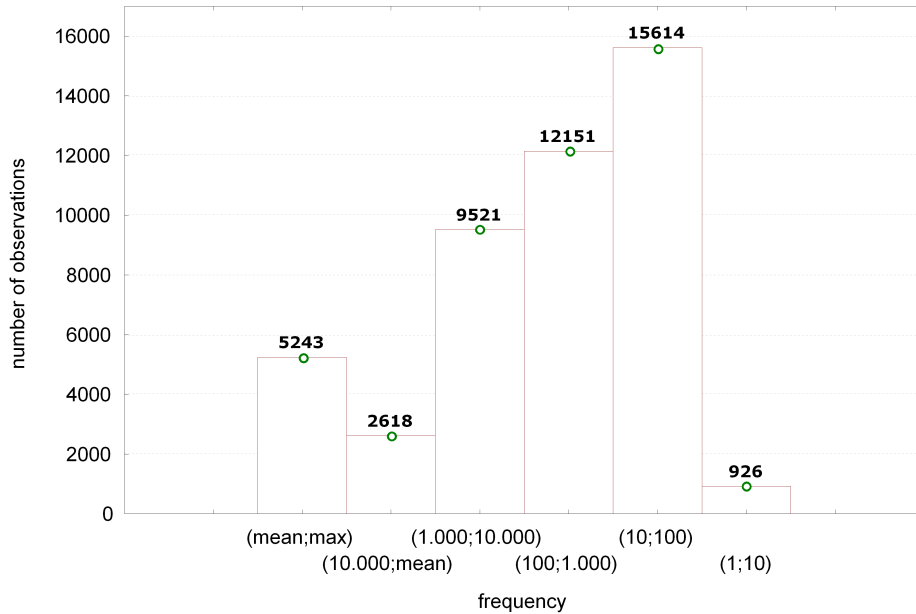


Figure 1: Distribution of frequency in the inflected dictionary list

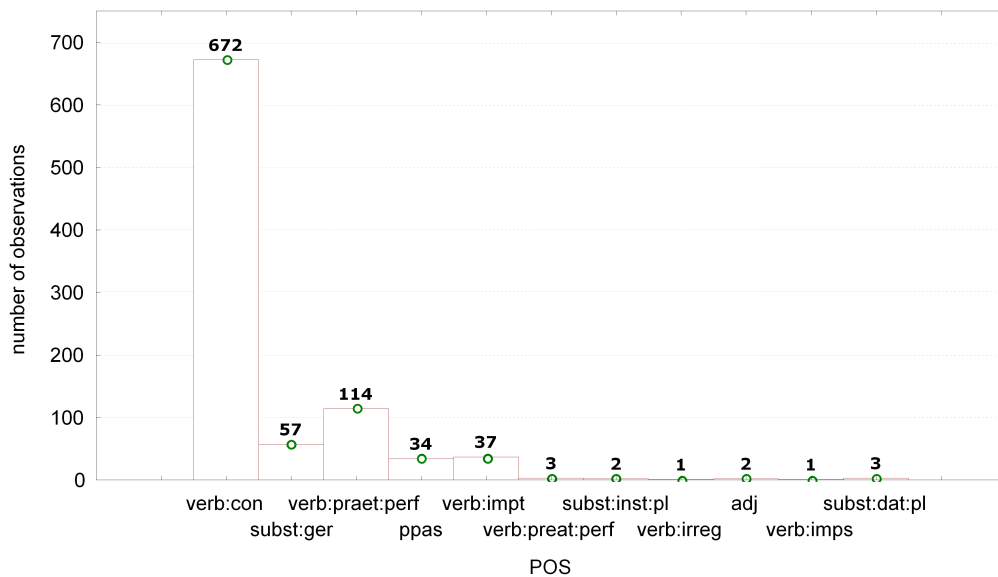


Figure 2: Distribution of POS in the least frequent ($1 < f < 10$) word group

The analysis has confirmed the usefulness of the pre-selection of the less frequent words. Moreover, some more rare grammatical forms have been found. Google search has shown that verbs in the conditional have very low frequency: they constitute the most numerous group of words with the lowest frequency ($1 < f < 10$) and many of them are singletons

($f=1$). This is illustrated in the Figure 2. Since our impression was that not all of them occur indeed so rarely and that the frequency depends on a number of factors (including the frequency of the verb in the infinitive), final decision on deletion/leaving of the word from the list was taken by a linguist.

Once all missing frequencies were found the procedure similar to that applied in the extraction of the common word list was used. So, the word frequency in each domain has been determined as follows:

6. The words were sorted with the decreasing frequency.
7. Singletons were removed
8. Relative frequencies $Rf(w)$ were calculated: the frequencies of all words were summed up $Sf(w)$ and then each observed word frequency $Of(w)$ were multiplied by 100 and divided by $Sf(w)$
9. self-covarege (SC) of the word list was determined by summing up $Rf(w)$, see equation (1)
10. the level at which the list should be truncated was determined: in the common word list this level was set at 96%. It was assumed that for the dictionary list the level would be similar.
11. Frequencies found in Google differ from those determined on the basis of the common word corpus (which can be seen when comparing the mean and median values), so we have to find a way to make the dictionary and CW list frequencies uniform (the work is in progress...).

6. From word lists to lexicon

Each lexicon entry is transcribed phonetically: an inventory of 39 phonemes was employed for the broad transcription and a set of 87 allophones was established for the narrow transcription of Polish [14]. The phonetic labeling was done automatically [17] and corrected manually. Syllable boundaries were marked according to [15] where a statistical approach based on the information concerning syllable structure and sonority profile is proposed. Stress was assigned according to the rules defined in [16]. In Polish the 'default' stress position is the penultimate syllable but it can also fall on the antepenultimate or ultimate syllable; in polysyllabic words secondary stress may fall on the initial syllable.

A preliminary POS tagging was based on [3]. But after testing a number of morphological analyzers available for Polish it was decided that the most useful information and highest accuracy can be obtained with [18].

7. References

- [1] Ziegenhain, U. et al. 2002. Specification of corpora and word lists in 12 languages. LC-STAR Deliverable D1.1.
- [2] Wierzchoń, P. 2007. Rules of building Polish text corpora for lexicon creation. Technical report
- [3] Ispell-pl. Polish dictionary for ispell. <http://sourceforge.net/projects/ispell-pl> visited 06-Mar-07
- [4] Piotr Muldner-Nieckowski. Wielki słownik skrótów i skrótowców. Europa, 2006
- [5] Jerzy Bralczyk. Słownik 100 tysięcy potrzebnych słów. Polish Scientific Publishers (PWN), 2007
- [6] Podręczny słownik medyczny. Wszystko o zdrowiu. Podstawowe terminy medyczne. Twój Styl, 2006
- [7] Jonathan P. Murray, Jerzy Radomski, Włodzimierz Szyszkowski. Język angielski w medycynie. English in Medical Practice. Wydawnictwo lekarskie PZWL, 2005
- [8] Ewa Rostek i Krzysztof Rostek. Uniwersalny słownik tematyczny języka niemieckiego. WAGROS, 2005
- [9] Słownik terminów komputerowych i informatycznych. Rea, 2006
- [10] I. Kurecz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak: Słownik frekwencyjny polszczyzny współczesnej [Frequency dictionary of contemporary Polish]. Instytut Języka Polskiego PAN, Kraków 1990, vol 1 i 2
- [11] Imiołczyk Janusz. Prawdopodobieństwo subiektywne wyrazów. Podstawowy słownik frekwencyjny języka polskiego [Subjective probability of words. Basic frequency dictionary of Polish]. Warszawa, Poznań: PWN, 1987
- [12] M. Bańko, M. Krajewska, Słownik wyrazów pospolitych, Warszawa (SWK), 1994
- [13] The PWN Corpus of the Polish Language. Polish Scientific Publishers, Warszawa 2007

- [14] Demenko, G., Wypych, M., Baranowska, E. 2003. Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis. In: Demenko, G., Karpiński, M. (eds.), *Speech and Language Technology*. 7, 79–97
- [15] Sledzinski, D. 2004. The numerical and sonority profile aspects of the Polish consonant clusters. In: Demenko, G., Karpiński, M. (eds.), *Speech and Language Technology*. 8, 65-74
- [16] Steffen-Batogowa, M. 1996. *Accentual structure of Polish*. Poznan: UAM
- [17] Szymanski M., Grochowski S. 2005. Transcription-based automatic segmentation of speech. *Archives of Control Sciences*. 15, 465–472
- [18]