How many words do you know in your native language?

The analysis I am conducting in this paper is an attempt to answer the question stated in the title. The language the analysis is based on is Polish, however, I will refer to English in the later part of the paper to provoke further discussion. My native language is used here as a tool to constitute a hypothesis. Nevertheless, to set a structure for the discussion on the issue and build up the categories for the analysis, I will focus on the description of 3 main components of the question in the title. Namely: "how many", "word(s)", "know". Defining those 3 crucial parameters will certainly affect the results of the analysis. Any conclusion drawn from the data examination will be dependent on the definition purposefully chosen and stated to set the boundaries for the analysis.

The first component ("how many") indicates that a fixed number is needed to be concluded from the analysis. To build an algorithm used in NLP a clear numerical structure and size of words is necessary to make computation possible. The algorithms working recursively and basing on repetitions need to have a number to be compared to a different dataset of a different size. Therefore, no ontologies or linguistic computational structures can be created if the numbers are not known: how many words a language consists of; how many words an average person is aware of, etc. The numerical data is needed also for the diachronic purposes and investigating the changes in a language structure, as well as for the comparatistic purposes - my native language vs English.

The second component ("word/s") can be described as a basic form of a word but not an individual morpheme. For the sake of my analysis, I do not acknowledge a single morpheme as a word or a word that is not in its basic form and is derived from the stem with the use of a morpheme. E.g. the suffix -ed is a linguistic unit that bears a meaning and constitutes new semantics added to the verb "look", i.e. look -> present tense, look-ed -> past tense. However, in the paper I support that the purpose of the linguistic analysis is to find out the knowledge about the stems of words, its basic forms that may be used for further derivation. Polish is a synthetic and fusional language, hence taking all forms of a stem into consideration may skew the data significantly and the number of tokens can increase very quickly. Additionally, before conducting the analysis I clean the data, use all the words lowercase and do not include symbols like emoticons, spaces, and punctuation marks. Although, I do consider function words.

The third component ("know") indicates in this analysis that I do not need to search for the accurate word to convey the meaning and express the message. Therefore, a word is considered known if as a native speaker of Polish I am able to explain the meaning of the word, even though the semantic representation may differ depending on the user's understanding. This point refers to the biggest problem of computing language - ambiguity. For the sake of this analysis, if a word's meaning can be explained by the user and the use makes sense in the context without looking it up, the word is "known".

The resources used in the analysis of the paper:

- As a numerical basis for the comparisons of my analysis, I will use the Polish dictionary "Wielki Słownik Języka Polskiego" ("The Great Dictionary of Polish Language", my translation; the dictionary available online, https://wsjp.pl/).
- To set the database for the words used in the analysis I exported 7 different private WhatsApp conversations and concatenated them into one text file. The purpose was to provide the real data and clearly support the argument that a word I use in my native language and consider as "known" does not need to be researched or looked up.
- In order to compare the analysis with the resources available online, I will use the AGH corpus of Polish speech and the paper written at the AGH University of Science and Technology, Kraków, Poland. ("AGH corpus of Polish speech", Piotr Zelasko • Bartosz Ziółko •

Tomasz Jadczyk • Dawid Skurzok, published online: 6 May 2015, https://link.springer.com/content/pdf/10.1007/s10579-015-9302-y.pdf)

The analysis of the real life data from WhatsApp and data cleaning process was conducted in Jupyter Notebook using Unix commands. To visualize the data and prepare the graphs I use Python.

After concatenating 7 text files and counting the words I have: 3222 lines, 35319 “words” and 25M bytes:

```
! wc chat.txt
```

```
  3222  35319 251551 chat.txt
```

Applying Heap’s Law with the parameters: k = 10, b = 0.6, the prediction for working vocabulary is: 5350 words.

After listing all alphabetic characters and counting the number of lines, I get the number of 5993 tokens:

```
! tr -sc 'A-Za-z' '\n' < chat.txt | sort| uniq | wc -l
```

```
5993
```

The following analysis changes all the words into lowercase characters and the number of words used by me as a native speaker of Polish is: 5270 (out of 35319).

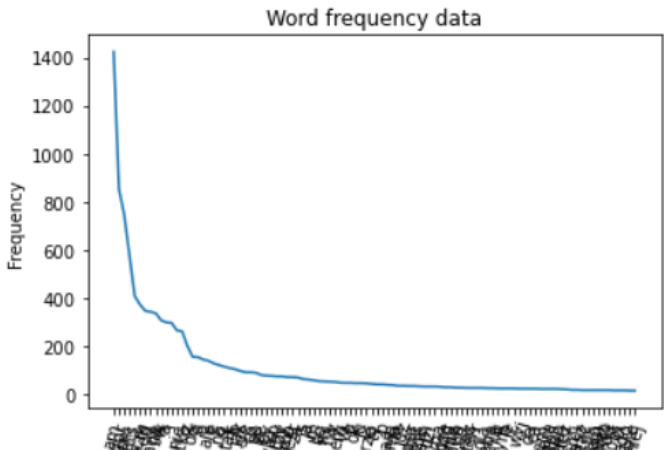
```
! tr 'A-Z' 'a-z' < chat.txt > chatlowercase.txt
! tr -sc 'a-z' '\n' < chatlowercase.txt | sort| uniq | wc -l
```

```
5270
```

To continue the analysis, I will compare the results with the numerical basis mentioned earlier in the paper, The Great Dictionary of Polish Language. The dictionary consists of approximately 50000 unique entries in their basic, not inflected, form (in 2018). Therefore, taking into consideration 7 different chats with roughly 5300 unique words, as a native speaker of Polish I use about 11% of the vocabulary range set by The Great Dictionary of Polish Language in everyday conversations.

As above mentioned, the third component setting categories for the analysis (the word “know”) indicates that to know a word means no need for looking it up in a dictionary or researching for the meaning that is easy to explain for the user. Therefore, I will compare the results with the corpus statistics from the paper “AGH corpus of Polish speech”. This kind of corpus presents the transcriptions of spoken Polish which is more similar to everyday conversations exported from my WhatsApp database. The AGH corpus contains 117450 words where 13724 words are unique. Compared to the number of unique words used in the AGH corpus, the words used in my chats constitute 38% of the AGH corpus words.

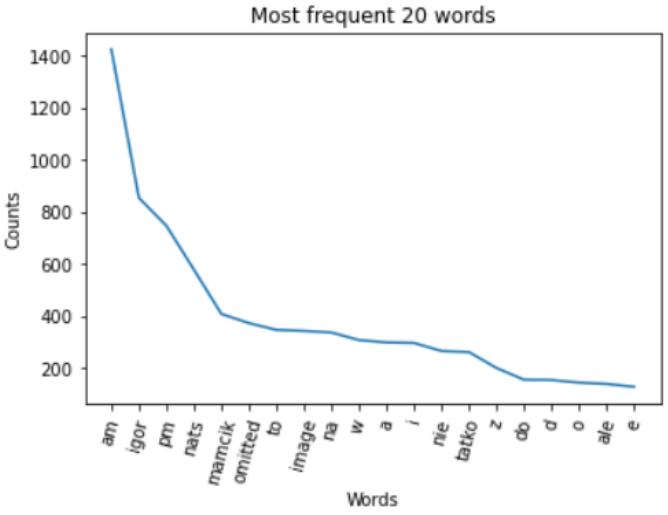
The frequency of the most common words is very high and drop drastically after the first most frequently used ones:



The interesting observation is that the most common words used in my private WhatsApp chats and presented in the AGH corpus statistics are similar:

Table 2 List of 60 most frequent words along with their translation to English and their occurrence frequency in our corpus and in a 1 billion words corpus from earlier works (Ziółko and Skurzok 2011)

Word	Translation	Occurrence in AGH corpus (%)	Occurrence in (Ziółko and Skurzok 2011; %)
na	on	2.55	1.67
w	in	2.38	2.26
z	with	1.89	1.51
się	*a*	1.54	2.39
i	and	1.48	2.34
do	to	1.30	1.10
nie	no	1.18	1.82
jest	is	0.94	0.43
o	at	0.90	0.54
to	this/it	0.87	0.98
poproszę	I'd like	0.83	0.00
jak	how	0.44	0.51
czy	*b*	0.41	0.24
co	what	0.37	0.40
włącz	turn on	0.37	0.00
bilet	ticket	0.36	0.00
dwa	two	0.35	0.05
że	that	0.34	0.95
po	after	0.33	0.44
a	and	0.33	0.67



The graph to the right is the plot of the first 20 most frequent words used in my conversations. The table to the left presents the first 20 most frequent words from the AGH corpus. Prepositions, pronouns, conjunctions and interjections are the most frequently used in both corpora and constitute most of the vocabulary range used in conversations. Referring to the definitions from the introduction and the main question stated in the title: are those real words used in Polish?

Additionally, a lot of English words or words derived from English are present in my dataset of WhatsApp chats. Therefore, another question arises if those are allowed to be counted in the Polish vocabulary. They do not originate from my native language, however, they are highly used in everyday conversations and convey the meaning of the messages.

In conclusion, focusing only on the numerical data I gathered is not accurate and enough. Comparing the number of types from my WhatsApp data to the Polish dictionary gives a mediocre number. However, what are the categories used by lexicographers to define the entry and justify the choice? The comparison to the AGH corpus gives a slightly more promising number, however, my data is skewed and focuses narrowly on specific subjects. The conversations come mostly from chats with my family so one of the most frequent words are our names. A crucial conclusion drawn from the analysis is that function words constitute most of the words practically used in the analyzed corpora and are mostly used by me as a native speaker of Polish.

Natalia Wojarnik