MASARYK
UNIVERSITY

FACULTY OF INFORMATICS

Age-of-Acquisition Ratings for Czech Words

Master's Thesis

DAN MAKALOUŠ

Brno, Spring 2022

MASARYK UNIVERSITY

FACULTY OF INFORMATICS

Age-of-Acquisition Ratings for Czech Words

Master's Thesis

DAN MAKALOUŠ

Advisor: doc. Mgr. Radek Pelánek, Ph.D.

Brno, Spring 2022


FACULTAS ARTIS INFORMATICÆ
UNIVERSITAS MASARYKIANA

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Dan Makalouš

Advisor: doc. Mgr. Radek Pelánek, Ph.D.

Acknowledgements

I would like to thank my advisor, doc. Mgr. Radek Pelánek, Ph.D., for his help and willingness to discuss problems I encountered. I would also like to thank the experiment participants, especially the children, who were essential to evaluating the result. And finally, great thanks belong to my friends who read the text a provided me valuable feedback, namely Přemysl Till, Adam Gregor, and Vojtěch Spěvák.

Abstract

This thesis presents age-of-acquisition ratings for 32 954 Czech words. It consists of gathering sources, assessing their quality, and combining them to create the resulting data set. These sources included word properties (length, frequency), foreign age-of-acquisition studies, word embedding vectors (to find and utilize similarity of words), and a self-conducted experiment where we collected 5 778 subjective age-of-acquisition estimations for Czech words. Because of the variety of used sources, the final estimates are of diverse quality, which we address by including certainty of the rating in the resulting data set. The expected use of the results is to estimate the difficulty level of words in texts and check if it fits the age of the expected reader.

Keywords

age-of-acquisition, estimation certainty, word embeddings, AoA collection

Contents

1 Introduction

Age-of-acquisition (AoA) of a word is the age at which more than 50% of the population can understand its meaning and use it. Its estimation could be utilized in numerous ways. One possible use, which is also the primary motivation for this thesis, is to provide a valuable tool for the authors of text comprehension exercises. Those are exercises that test how well a person (usually a child or a student) can understand a content of written text and how capable they are of finding information in it. The text author will be able to use the data set to check if the AoA of individual words is adequate to the expected age of the child/student. Another possible application would be to use the estimations to presort words by AoA or to select the most difficult or the easiest words in a list.

It has been shown that the AoA of words is an excellent estimator of word difficulty, even more so than word frequency or syntactic complexity [1]. It also affects the lexical processing of words, for example words with lower AoA are thought to be read faster [2]. Frequency or length of words are often controlled for in psycholinguistic studies, and there is a strong case that controlling for AoA is even more critical.

Since 2012, several studies have created data sets covering the most used vocabulary in a language with AoA ratings [3, 4, 5]. Every such study known to us relies on AoA estimates from adults, so-called **subjective AoA**. However, the best way to gather AoA of words is to ask children of different ages if they know and can use various words; that would be collecting the **objective AoA**. Although intuitively, the subjective AoA does not seem viable (because adults do not truly know when they have acquired words), the studies on this topic show that their guesses are generally highly accurate. The study by Morrison, Chappel and Ellis in 1997 [6] was one of the first studies to report that the correlation is about 0.75 by comparing subjective and objective estimations. Since then, there have been many validation studies confirming this claim, e.g. [7, 8].

In spite of the benefits of having AoA ratings, there is very little knowledge about the AoA of Czech words. Therefore, this thesis aims to estimate the AoA for a large number of Czech words (>30 000). Because existing resources are for the Czech language very limited,

we aspire to create only rough estimations which could serve more as a helpful tool than a reliable metric. However, together with every estimation, we also aim to determine the estimation certainty, which should indicate how accurate and trustworthy each estimation is. That would allow filtering of estimates by their accuracy and thus expand the possible uses.

The approach of this thesis is to create an experiment collecting Czech AoA for words (Chapter 3), then gather sources (frequency table, English ratings, Dutch ratings) and estimators (word length, frequency) which could be relevant to Czech AoA, evaluate and combine them (Chapter 4). The relevancy of sources is determined by their comparison to collected data, which we use as ground truth. The evaluation in Chapter 5 shows both the accuracy of the final estimates and the relative comparison of individual estimators. To demonstrate the utility of our results, we also included evaluations by objective AoA (data from children). The final chapter discusses the accuracy of the final estimates, the possible uses of the created data, and lastly, the limitations of our approach and possibilities of future research.

2 State of the Art

Age-of-acquisition has been studied as a psycholinguistic variable since the middle of the previous century. Still, it has only been a decade since since the first appearance of large AoA data. Those opened up other research directions, such as studying their relation to other (more available) variables or prediction of AoA via machine learning methods. The following sections will discuss the three main directions of research.

2.1 Questionnaire Studies

The most successful way to obtain AoA for many words thus far is a comprehensive questionnaire. People are asked to estimate the age when the word is usually acquired. The idea of such studies is to collect as many answers for every word as possible and average them. To accurately estimate the most used words in a language by this method (tens of thousands of words), it is necessary to gather hundreds of thousands of answers.

The first large-scale experiment of this kind was done by Kuperman, Gonzales, and Brysbaert in 2012 when they collected AoAs for more than 30 000 English words [3]. They used Amazon Mechanical Turk for collecting data and paid each participant a small amount of money for every labeled word. In this manner, they collected a total of 842 438 ratings. The data set was later enlarged from 30 000 to 44 000 words by Brysbaert and Biemiller [9]. Brysbaert has further participated in another study from 2014 also used in this thesis, where they gathered AoA estimations for more than 30 000 Dutch words [4].

2.2 Word Features

Another approach for estimating AoA is to use other variables correlated to AoA. Usually, the focus of such studies is to determine the properties of AoA more so than to estimate it for a large number of words. However, in the case of this thesis, when the aim is only to generate rough estimates, using less precise but more available sources

can be a useful technique. The most studied variables concerning AoA are these:

- **Length** of the word intuitively relates to AoA. The easiest words in a language typically have only one or two syllables, and the longer the word is, the higher the AoA. The most commonly used length property is the number of letters; there are also studies using the number of phonemes or syllables (but the choice does not change the outcome significantly). The correlations to AoA are usually between 0.3 and 0.5 [5, 10, 11]. The important advantage of word length as an estimator is that it is very available (can be directly derived from word), which is very useful in our case.

- **Frequency** of the word is determined by the number of occurrences of the word in a large corpus. That might often be an issue, because, for example, elementary words commonly appearing in children's books are often not very frequent in the used corpora (that is also the case of this thesis). The more frequent a word is in a language, the lower its AoA. Because frequency usually has a very broad range of values, AoA is usually predicted using its logarithm. The found correlation across different studies is in the range of -0.4 to -0.6, and in most cases, frequency is a much better estimator of AoA than the word length. After word length, word frequency is typically the most accessible variable, and it is available for a wide range of languages, including Czech [8, 12, 13].

- Although the AoA of words is language-dependent, **AoA across languages** are often highly correlated. In an article from 2019, Łuniewska, Wodniecka et al. [11] compared the ages of acquisition of 300 words from various languages, including Czech. They used available data from other studies and did an experiment for seven new languages. According to this study, the Spearman correlation among different languages is between 0.49 and 0.84. It also shows that languages from the same region or the same language family tend to have similar AoAs.

  For us, the relevant correlations are Czech to languages that have large, openly available data sets with AoA ratings. These

are English and Dutch, and the correlations are 0.77 and 0.85, respectively. The article also reports that Czech AoA is positively correlated to word length by 0.33 and negatively to frequency by -0.45.

- Other variables known to relate to AoA are **Imageability, Familiarity, Concreteness**. However, though they are excellent estimators of AoA, they are not available for the Czech language and therefore are not relevant to this thesis. Imageability represents how well the word invokes an image, concreteness indicates the degree of abstractness, and familiarity refers to how well-known the word is in the language. The imageability negatively correlates with AoA in rates from -0.3 to -0.6, and it also highly correlates with concreteness [14]. Familiarity relates to AoA very closely. A Portuguese study from 2010 [15] found a positive correlation of 0.68.

An example of the approach using word features related to AoA is an Italian study by Russo [10]. He used a conjunction of multiple data sets resulting in AoA for 1 908 Italian words. They estimated 2 783 lemmas using linear regression of word length, frequency, the logarithm of frequency, concreteness, and imageability. The regression on all those features correlated to AoA by a factor of 0.379, and a combination of word length and the natural logarithm of frequency by 0.378.

2.3 Word Embeddings

Another way of predicting AoA is to use word embeddings. Word embeddings represent words using vectors of values (typically using hundreds of dimensions). They are obtained by analyzing contexts of words in large corpora. When using word embeddings, the primary objective is to grasp the semantics of words and thus determine which words are similar in meaning and to what extent. A popular tool for this is the word2vec model [16]. Methods which use word embeddings are often referred to as corpus methods. That is because the corpus used for training the embeddings determines their nature.

When trying to obtain the embeddings for the whole language, it's essential to choose a corpus that is a good representation of it.

Mandera [17] studied how effective corpus methods are in extrapolating psycholinguistic variables, AoA included. The best performing corpus model was KNN (K-Nearest-Neighbors), with a correlation of 0.74. For comparison, a linear model with frequency logarithm as the only estimator correlated with AoA by 0.62.

Botarleanu et al, [18] simulated the process of learning words by training word2vec on corpora that were incrementally increasing in size with every epoch of training. This study was done on multilingual texts. By simulating a person learning words, they could recognize features crucial in word acquisition and compare AoA of different languages. They were able to determine learning phases of English, German, French and Spanish and for example found that Spanish language has on average lower AoA than the other languages.

3 Data Collection

An essential part of this thesis is an experiment where subjective estimations of AoA for Czech words are collected. The experiment is a simple web page where people communicate with a system estimating AoA of words. The motivation for doing this experiment is twofold: firstly, gathered data can be used directly as an AoA estimator, and secondly, data can be used as ground truth to evaluate the accuracy of other methods (translations, frequency, word length).

The population for this experiment was not controlled. All the participants took part voluntarily, anonymously, and were not paid. The experiment was distributed mainly through academic groups and forums and snowball sampling.

After accessing the page, the system generates a word for the user to label. After the user inputs their subjective estimate about the word and hits "send" (or presses enter), the system provides feedback by revealing its estimate in a table below and generates another word to label. This process repeats indefinitely; there is no fixed number of words a user should label. At the bottom of the screen, the user is informed about the purpose of this experiment and that it is ongoing i.e. that they can label as many words as they want.

3.1 Requirements

- **Cover as many words as possible.** The most successful studies covered all words in their data sets with multiple answers. That is not an option in our case due to the limited scale of the experiment and available resources. For the results to be usefully applicable, we need to collect a non-negligible number of them. We estimated at least 5% of the whole data set, about 1500 words.
- **For a subset of words, collect multiple answers.** To be later able to evaluate the accuracy and relevancy of other estimators, it is very useful to have a set of words for which we are reasonably certain about their AoA.
- **Make it interesting for the user.** The participants of the experiment are not paid, and the spread of the experiment is limited. It is

therefore not expected to reach an enormous number of people, and it's important that the ones it reaches enjoy the participation, so they hopefully provide answers for more words. To achieve this, we covered these points:

- Give appropriate feedback. We hoped that by giving participants feedback, they would also be curious when the words are acquired and therefore motivated to continue the experiment.
- Don't ask a user to estimate the same word more than once. That is both because it is bothersome and because the participant could remember the provided feedback and respond with that (which would bring bias into the data collection).
- Generate words with different certainties. We want to gather new information about words we are not certain about, but we also wish to be able to provide the user with relevant feedback (at least sometimes). A repeated selection of already estimated words (by other users) allows to provide accurate feedback and additionally causes a set of words to be evaluated by more answers, which is utilized in evaluation.

3.2 Feedback Design

The system gives feedback via the table shown in Figure 3.1, where a user can see his last 20 answers and their comparison with the system's estimate. There are five columns:

- Slovo (Word) — the estimated word.
- Uživatel (User) — the user's AoA estimation of the word.
- Comparison column — comparison of the user's and system's estimation, using following symbols {<, >, ~}. The ~ appears when the system and user differ by 1 year at most.
- Systém (System) — current system's AoA estimation of the word. The saturation (opacity) of the estimation depends on the system's estimation certainty.

Figure 3.1: The experiment design. A user estimates the red-colored word. The table shows the user's and system's estimations for previous words.

- Jistota systému (Certainty of the system) — the system's certainty of estimation. There are 6 different values ranging from very high to very low.

Aside from the fact that giving feedback makes participating in the experiment more interesting, it can be helpful when a user repeatedly rates words too high or too low. They can adjust their estimations by comparing them to the system's estimations using the feedback table 3.1. This effect is enlarged by the color differentiation of comparison operators.

To give users an anchor from which they can derive their estimations, we prepared a calibration table (available in attachment B). The table contains AoA derived in a small-scale experiment by collecting the objective AoA (asking children) for 57 words (a noun, an adjective, and a verb for every year from 2 to 18). Because these words were excluded from the pool from which we selected words to label, they can also be used to evaluate final estimations in Chapter 5.

3.3 System Evaluation and Certainty

The system certainty of AoA estimation is a discrete value from 0 (very low) to 5 (very high). The initial rating and certainty were based on data from Kupperman's study [3]. To words for which we were able to find AoA based on this study via translations (8549 words), we assigned the certainty of 2, for the rest we used the closest words estimation discussed in 4.2.4 and assigned the certainty of 1.

When users and the system agree on the AoA of a word, the certainty of the system's estimation for the word increases; if they disagree, the certainty remains or decreases (based on the difference). The system updates its AoA estimate, so it is always the average of all users' estimates (answers_AoA) and the initial estimation (initial_AoA) of the word given by the formula below. The initial AoA rating is counted as 3 answers from participants so that 1 response from the user does not change the estimate completely and only sways it in the right direction.

$$AoA = \frac{3 \cdot initial\_AoA + \text{sum}(answers\_AoA)}{3 + \text{count}(answers\_AoA)}$$

3.4 Implementation

The experiment is available at `https://www.fi.muni.cz/~xmakal/AoA`. It runs on the FI MUNI server "Aisa" as a CGI script. A CGI script was chosen because it is a quick solution, and the scale of this experiment is relatively small.

The solution does not use cookies, the necessary information is stored as URL parameters (id of the word, id of the user, estimation). The user's id is a randomly generated hexadecimal number assigned when they access the web page. It is then passed as a URL parameter and stored in logs, so it is possible to find the user's passed estimations. This solution also does not use any database system, and all the data are stored in logs and in CSV file. Technical details are omitted because they are mostly ad hoc solutions not important for the thesis.

3.4.1 Word Selection

Selecting which word will the user label influences both the nature of collected data and how enjoyable the labeling will be. It is largely random. We filter out all the words, which the user has already estimated. Then a word is selected as follows:

- With a 20% probability, it selects one of the 20 words with the highest certainty to give the user reliable feedback and cause some words to have many answers.
- With a 30% probability, it selects one of the 30 words with the lowest certainty. This mostly works the same as randomly choosing from words which the user has not already estimated. That is because this experiment will not cover the whole data set, and the majority of the words will have the initial certainty of 1 for the entire duration of the experiment. It can, of course, happen that a word drops to certainty of 0, and then this choice prefers it.
- Otherwise, a random word from the filtered word set is selected.

3.5 Collected data

The experiment resulted in 5 778 separate answers from 294 sessions. On average, a user labeled 19 words in one session. Twelve answers were filtered out because they disagreed with other answers and our estimation to such an extent that we considered them to be mistakes. After grouping the answers by words, we have 3 517 labeled words, from which 17 were filtered out because of high deviation between answers. Both the filtered logs and answers can be found in attachment A.3. That leaves us with 3 500 words labeled with Czech AoA.

Figure 3.2 shows for how many words we have a certain amount of answers. We have only one answer for 2 586 words, which gives us 914 words with multiple answers. This satisfies the requirement that we should collect data for as many words as possible but also have a good amount of words with multiple answers for evaluation purposes.

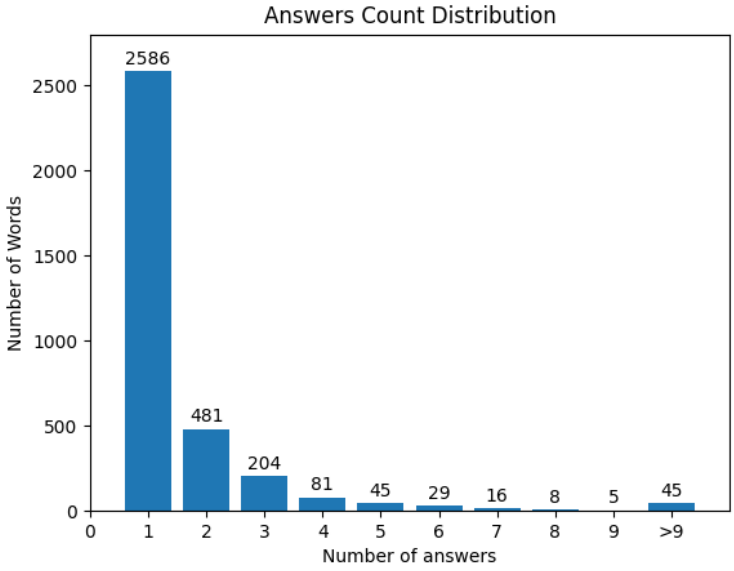

Figure 3.2: Number of answers for words in the experiment.

Figure 3.3 displays the AoA distribution of collected data. Had the methodology of this experiment been ideal, this distribution would show how many words are learned at what age proportionally. We can see that the distribution is not uniform; the number of learned words rapidly increases up to the age of 9 and then slowly decreases. The fact that the regression has a clear peak will be relevant when creating an estimator of Czech AoA.

To check the internal consistency of collected data, we used the 914 words with multiple answers to do a split-half reliability test. Every word with more than one answer had its answers divided into two halves. We then measured the correlation between the means of the halves shown in the Figure 3.4. The correlation was 0.72, which shows that collected data is internally consistent and therefore a good predictor of subjective AoA.

The stability of the prediction depends on the number of answers used to create the estimation. That is demonstrated in Figure 3.5. With an increasing number of answers, we can see that error rapidly decreases exponentially with respect to the number of answers.

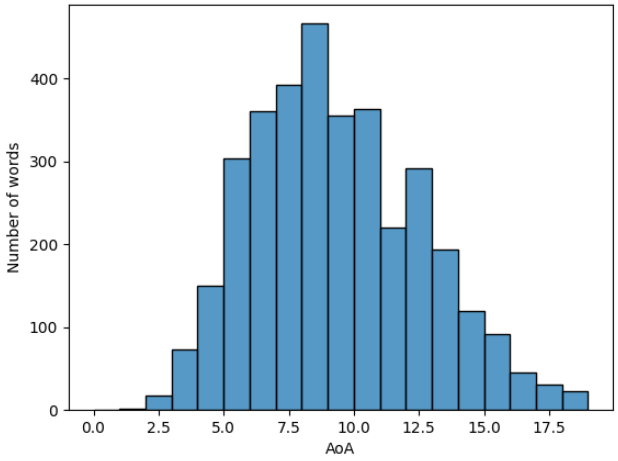

Figure 3.3: Distribution of AoA in collected data.

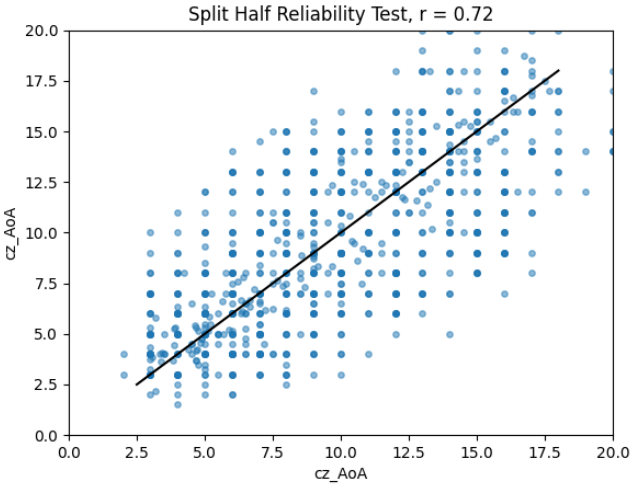

Figure 3.4: Split half reliability test. Compares two different estimations by splitting answers for every word into two halves. The data points are transparent, darker color means there are more points on top of each other.

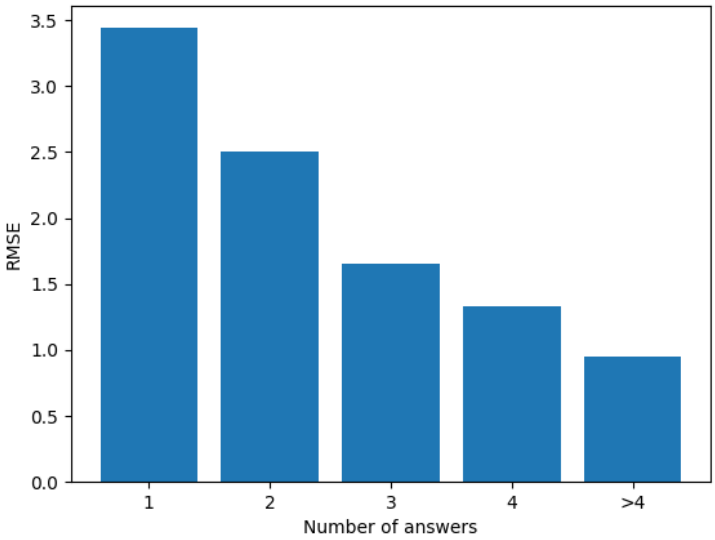

Figure 3.5: Root Mean Squared Error between the two halves depending on the number of answers used for estimation.

4 Method

The approach of this thesis is to find, evaluate and combine various sources and techniques for estimation of the words' AoA. Diagram 4.1 depicts the whole process, which sources were used, and how they were combined. In summary, the steps are:

- Czech words, Filter — It's important to choose a reasonable word set to which we will try to assign AoA. It should contain all the words that could appear in common Czech publications, books, or generally in everyday speech. We then filter this data set by choosing only the most frequent 30 000 to 40 000 words and further removing nonsensical, inappropriate, or other unwanted words.

- English AoA, Dutch AoA, Translate & Map — We use large-scale foreign studies (English, Dutch). We translate the words and map their AoA onto the Czech counterparts in our data set. By this process, some information will be lost because the translations will be imperfect, and the AoA will differ in other languages.

- Collect Czech data — The collection of Czech data was discussed in Chapter 3. Because the collection requires the system to have its own AoA estimation for every word, it is necessary to make first rough estimations based on the available data.

- Word Set, Czech AoA, Evaluate + Transform — We have collected Czech AoA for about 10% of the whole data set and it is (as a subjective AoA) the best source of AoA from what we have. It was used to evaluate the quality of other estimators (frequency, length, translations, word embeddings) and transform them by (linear) regression.

- English Estimation, Dutch Estimation, Word Embeddings, Combination — Combining AoA estimators by weighted average based on their performance. That results in the final estimator, which was used to create the final data set of AoA estimations.

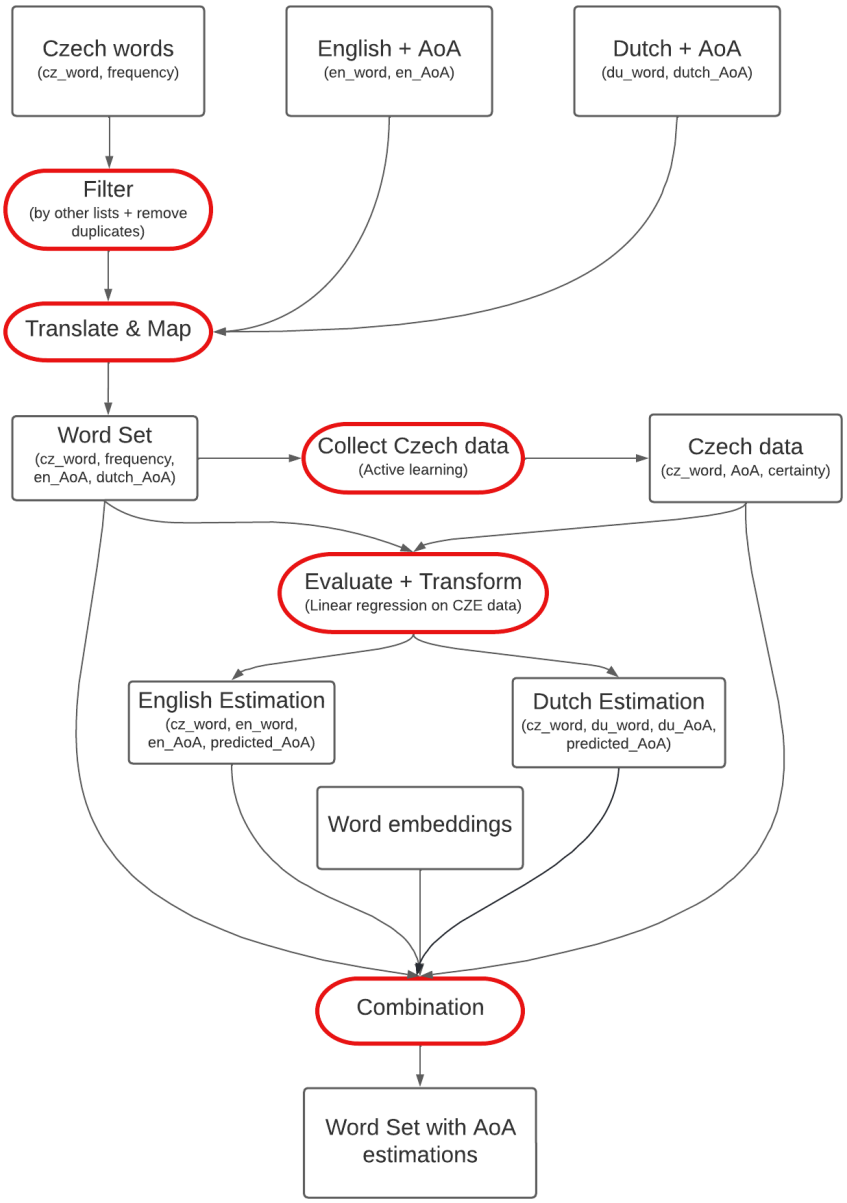

Figure 4.1: Flow diagram. The black squares represent data; the red ovals represent transformations (filtering, combining).

Table 4.1: Sources. The names are in capital letters to be easily recognizable in text. The column Useful data refers to used columns from the source.

Name	Description	Size	Useful data	Source
SYN2015	Standard Czech corpus with word frequencies	126 973	cz_word, cz_freq	[19]
LEX	Czech lexical corpus.	275 780	cz_word	[20]
ENGLISH	English AoA data set.	51 715	en_word, en_AoA	[21]
DUTCH	Dutch AoA data set.	31 179	du_word, en_word, du_AoA	[4]
CZECH	Data collected in our experiment.	3 500	cz_word, cz_AoA	Experiment
WORD SET	All data sources combined.	32 954	cz_word, cz_freq, cz_AoA, en_AoA, du_AoA	All of the above

4.1 Data Sources

One of our primary challenges in creating the AoA data set was finding and utilizing various data sources. The Table 4.1 is an overview of all the sources used to create the AoA estimators.

4.1.1 Foreign Sources

From foreign sources, the results of large-scale English and Dutch studies were used. English data set [21] is the largest openly available source for English AoA, broadening the data set from the Brysbaert

study [9]. The Dutch data set contains 31 179 lemmas and notably also English translations which partly explains the later discussed overlap of successfully mapped English and Dutch words. Analyzed were also AoA data sets from Polish [5], Italian [10], and German [22] studies, but because of their insufficient size or poor correlation to our collected data, we ultimately decided not to use them.

In order to use the AoA of words from foreign languages, we need to translate every word and map it to its counterpart in the other language. Because, to our knowledge, there is no openly available list of translated lemmas or dictionaries, we used a Python tool, PyPI, which uses Google translator. These translations are often not precise, especially for words with multiple meanings. Because of this, every word was translated from Czech to another language (English and Dutch) and back. If the translated word matched the original, we used its assigned AoA rating. The advantage of this technique is that we can be reasonably certain about the translation quality; on the other hand, we might be using only a fraction of available information. We matched 8 549 of the 51 715 English words and 6 011 of the 31 179 Dutch words.

4.1.2 Word Embeddings Source

To use corpus-based methods, we need to acquire the word embedding vectors. Training our own word embedding model based on some large Czech corpus would be very time-consuming, so we decided to use FastText.

FastText is an open-source library providing vectors for 157 languages, including Czech. It was trained on texts from Czech Wikipedia and Common Crawl. The advantage over other models is that FastText is able to generate vectors for words not included in its word dictionary based on word morphology [23].

With word embedding vectors for every word, we will be able to determine how similar any two words are by the cosine similarity of their vectors. This topic is further examined in section 4.2.4.

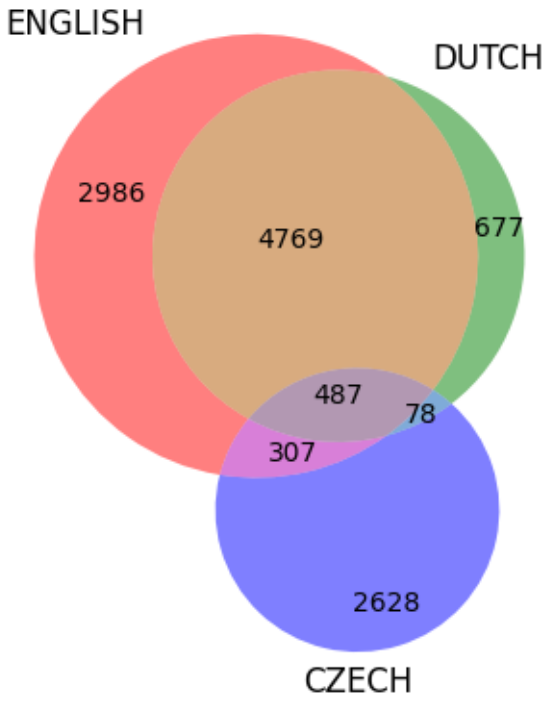

Figure 4.2: Source overlap within the WORD SET. A display of their comparative sizes and their overlaps. Together the sources cover 11 932 words.

4.1.3 Word Set

The WORD SET is a data set of 32 954 Czech words to which we will try to assign Czech AoA. For every word, it contains all available information (frequency, translations, closest words).

The primary source of Czech words was the standard frequency data set SYN2015 [19]. SYN2015 is the most popular source of Czech lemmas containing almost 800 000 of them, which practically covers all of the Czech language. The disadvantage of this source is that it was created by analyzing literature and newspaper articles. Therefore the reported word frequency might not reflect the word frequency of common speech, which would be more relevant for AoA.

The set goal was to rate about 30 000 most common words. Firstly, we selected the 40 000 most frequent words from SYN2015. These 40 000 also included strange words, e.g. common names. To eliminate those, we used the Czech lexical corpus (LEX) and filtered out words which were not included there (6 916). After that, we filtered swear words[1] (54) and one-letter words (77). That resulted in 32 954 Czech words with frequencies.

Then we used results from foreign studies (ENGLISH, DUTCH) and collected CZECH in the experiment. The composition of sources is displayed in Figure 4.2. We can see that ENGLISH and DUTCH almost completely overlap because the DUTCH data set also includes English translations, and thus it was very easy to find English AoA as well. CZECH overlaps with other sets minimally because the selection of which word was going to be labeled in our experiment was made pseudo-randomly from the whole WORD SET. Together, all the sources cover only 11 932 words, which is merely more than a third of the WORD SET we aim to estimate.

4.2 AoA Estimators

The sources from the previous section were used to create estimators of Czech AoA displayed in Table 4.2. These estimators vary in quality and can be combined to create better estimators or to cover more

1. For filtering swear words we used a list of Czech vulgarities from Wiktionary [24].

Table 4.2: The list of all estimators. Column Source refers to where the estimator came from, values Computed or Combined mean that it was created from other estimators. The last column #Words indicates for how many words the estimator is applicable (e.g. for estimator *du_freq*, it is necessary to have a Dutch translation).

Estimator	Description	Source	#Words
length	Length of Czech word.	SYN2015	32 954
freq	Number of occurrences in SYN2015.	SYN2015	32 954
log_freq	Natural logarithm of *freq*.	*freq*	32 954
en_AoA	AoA from ENGLISH.	ENGLISH	8 549
du_AoA	AoA from DUTCH.	DUTCH	6 011
cz_AoA	AoA from CZE_DATA.	CZE_DATA	3 500
cz_AoA1	cz_AoA with 1 answer.	CZE_DATA	2 585
cz_AoA2	cz_AoA with 2 answers.	CZE_DATA	484
cz_AoA3+	cz_AoA with >= 3 answers.	CZE_DATA	431
reg(*length*)	$0.59 \cdot length + 5.53$	Computed	32 954
reg(*log_freq*)	$-1.6 \cdot log\_freq + 20$	Computed	32 954
reg(*en_AoA*)	$12 \cdot en\_AoA - 0.35$	Computed	8 549
reg(*du_AoA*)	$1.28 \cdot du\_AoA - 3.1$	Computed	6 011
en_freq	$0.95 \cdot en\_AoA - 0.12 \cdot log\_freq + 0.67$	Computed	8 549
du_freq	$18 \cdot du\_AoA - 0.37 \cdot log\_freq + 1.3$	Computed	6 011
direct_AoA	Combination of *du_freq*, *en_freq*, *cz_AoA* by weighted average.	Combined	11 932
closest_AoA	Weighted average of 10 closest words with known *direct_AoA*.	Combined	32 954
final_AoA	The combination of *direct_AoA* and *closest_AoA*.	Combined	32 954

words. The final estimator covers the whole WORD SET and should be the best combination of all available estimators. It is a combination of *direct_AoA*, which uses all available data for the word (frequency, translations), and *closest_AoA*, which is the average of *direct_AoA* of the five semantically most similar words.

4.2.1 Atomic Estimators

The first type of estimators are atomic estimators. These are *length*, *freq*, *du_AoA*, *en_AoA*, *cz_AoA*. The estimator *cz_AoA* is further divided into *cz_AoA1*, *cz_AoA2*, and *cz_AoA3+* because of the difference in their quality[2]. All atomic estimators are directly taken from one of the sources in the Source table 4.1.

4.2.2 Computed Estimators

The computed estimators are *reg*(*length*), *reg*(*log_freq*), *reg*(*en_AoA*), *reg*(*du_AoA*), *en_freq* and *du_freq*. They are the results of a combination and transformation of one or more atomic estimators. The coefficients (weights) for combination were found on sub(sampled) data set by linear regression on *cz_AoA*. The data set was sampled because the age distribution in *cz_AoA* is not uniform and tends towards its mean. Sampling ensures that the resulting estimator does not prefer any age group.

For the final estimator, only *en_freq* and *du_freq* were used. They include the logarithm of frequency and originally also included the word length, but the regression coefficient was very close to 0, so we decided to exclude it completely. The other computed estimators (other than *en_freq* and *du_freq*) are included only to find their relation to Czech AoA.

4.2.3 Direct Estimation

Direct Estimation is a process of combining estimators for the word resulting in *direct_AoA*. It is called direct because it directly uses information available for the word (contrary to the Closest Words Estima-

2. The division of *cz_AoA* is made by the number of answers used to create the estimation (see the source table 4.1).

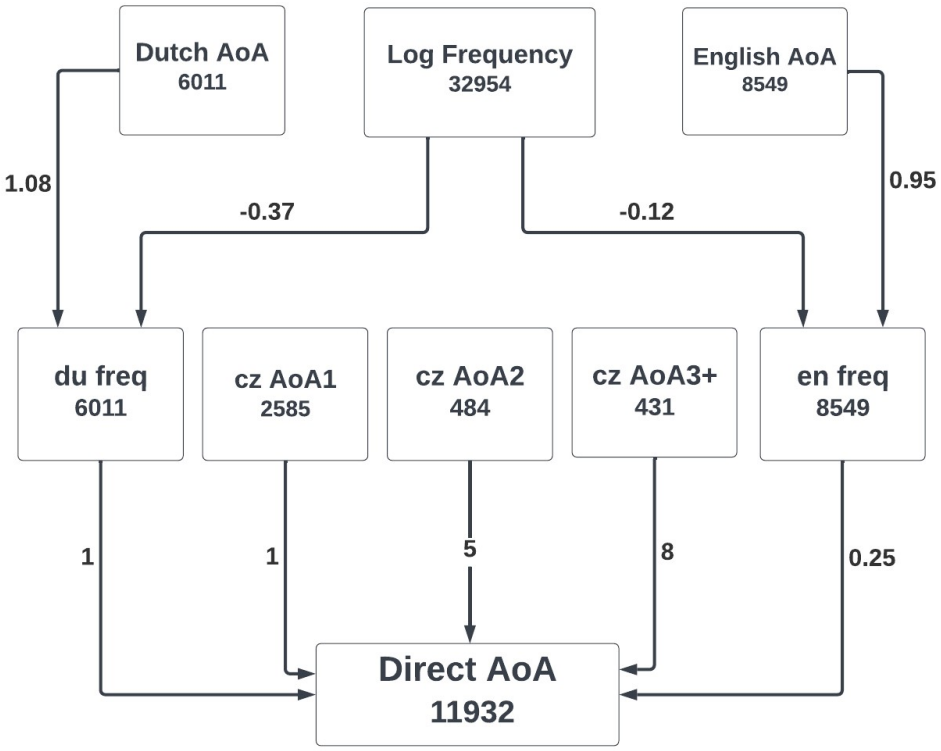

Figure 4.3: Direct Estimation. Every square box represents an estimator. Below the name of every estimator is the number of words the estimator is applicable on. By the arrows are weights, which are used when creating a new estimator using a weighted average.

tion, which uses information available for semantically similar words). The best performing estimators are *cz_AoA*, *en_freq*, and *du_freq*. How exactly the estimators were compared and how their accuracy was computed is described later in section 5.2. The *direct_AoA* cannot be computed for every word, because for most of them, we do not have any of these three estimators. For how many words is the estimator available can be read out of the Venn diagram in Figure 4.2. The ENGLISH corresponds to *en_freq*, DUTCH corresponds to *du_freq*, and CZECH to *cz_AoA*. By looking at the diagram, it's clear that the correct combination of *en_freq* and *du_freq* is crucial regarding the accuracy of Direct Estimation. Conversely, the other combinations are not that important because they only include a few words.

The overview of *direct_AoA* creation is depicted in Figure 4.3. There are three levels of estimators. The first level consists of Dutch AoA (*du_AoA*), Log Frequency (*log_freq*), and English AoA (*en_AoA*). These estimators were only used to create *du_freq* and *en_freq*, which, with *cz_AoA*, create the second level and together combine into *direct_AoA*.

The weights of the estimators were determined by regression on *cz_AoA*. Because *du_freq* and *cz_AoA1* are similar in quality and have little overlap in WORD SET, they are both set to weight 1 without more rigorous analysis. Regression coefficients for data with both English and Dutch translations resulting in the smallest error are in the ratio of $1 : 4$, respectively. Therefore the *en_freq* was assigned a weight of 0.25. The differences between estimations by *cz_AoA* with a different number of answers (*cz_AoA1*, *cz_AoA2*, *cz_AoA3+*) were already discussed and are compared by the bar chart in Figure 3.5. Based on that, *cz_AoA2* is given the weight of 5 and *cz_AoA3+* the weight of 8.

Apart from the creation of *direct_AoA*, the weights are also used to determine the accuracy of the estimation. Intuitively, the estimation created by the combination of three estimators is better than the estimation from only one estimator. The certainty of *direct_AoA* (*direct_certainty*) is the sum of weights of used estimators. For example if we use weighted average of *du_freq* and *cz_AoA2*, *direct_AoA* is $(1 \cdot du\_freq + 5 \cdot cz\_AoA2)/6$ and *direct_certainty* is 6.

Table 4.3: Closest words example.

klobouk			
close word	similarity	*direct_AoA*	*direct_certainty*
baret	0.64	13.4	1.25
šátek	0.57	4.64	1
deštník	0.51	4.9	1
kabát	0.51	2.49	1.25
turban	0.49	9.74	0.25

4.2.4 Closest Words Estimation

The main reason for computing *closest_AoA* is that *direct_AoA* covers only about a third of the WORD SET, but we need to estimate all the words. The evaluation by closest words uses word embedding vectors to find 5 of the semantically most similar words with known *direct_AoA*. Table 4.3 shows an example of such words for the word "klobouk".

Closest_AoA is the weighted average of *direct_AoA* of the 5 semantically closest words to the estimated one. There are 11 932 words with *direct_AoA*, and the closest words are always searched only within this subset. The word weight in the weighted average depends on its semantic similarity to the estimated word and the *direct_certainty* of its own estimation. The semantic similarity is determined by the cosine similarity of the words' vectors.

Closest_certainty is an average of $certainty \cdot similarity$ of the 5 closest words. The similarity is included because we want to give higher *closest_certainty* when estimating by more similar words. Note that because similarity is a number between 0 and 1, *closest_certainty* is always lower than the average of *direct_certainty* of closest words.

The weighted averages are expressed in the formulae 4.1 and 4.2. The sum over 5 closest words is denoted as $w \in closest$, $direct\_AoA$ as AoA and *direct_certainty* as $certainty$. The properties are accessed by an object notation (e.g. $direct\_AoA$ of a close word is $w.AoA$ in the formula).

$$closest\_AoA = \frac{\sum_{w \in closest} w.AoA \cdot w.certainty \cdot w.similarity}{\sum_{w \in closest} w.certainty \cdot w.similarity} \quad (4.1)$$

$$closest\_certainty = \frac{\sum_{w \in closest} w.certainty \cdot w.similarity}{\text{count}(closest)} \tag{4.2}$$

4.2.5 Final AoA

What is left is to combine *direct_AoA* and *closest_AoA*. The whole composition of the final estimator is displayed in Figure 4.4. The estimator ***final_AoA*** is the weighted average of *direct_AoA* and *closest_AoA* with the weights $1.5 \cdot direct\_certainty$ and $1 \cdot closest\_certainty$, respectively. And lastly, ***final_certainty*** is a weighted sum of *direct_certainty* and *closest_certainty* with weights 1.5 and 1, respectively. The weighted average is used only if the *direct_AoA* is available for the word. If it is not, $final\_AoA = closest\_AoA$ and $final\_certainty = closest\_certainty$.

For clarity, the weighted average and weighted sum are expressed in the formulae 4.3 and 4.4. In the formulae the word "closest" is shortened to "cl" and the word "direct" to "d".

$$final\_AoA = \frac{cl\_AoA \cdot cl\_certainty + 1.5 \cdot d\_AoA \cdot d\_certainty}{cl\_certainty + 1.5 \cdot d\_certainty} \tag{4.3}$$

$$final\_certainty = cl\_certainty + 1.5 \cdot d\_certainty \tag{4.4}$$

The correlation between mean absolute error and *final_certainty* is -0.12. That means that certainty predicts the quality of estimation but not very well. Moreover, the certainty is currently a number from 0 to somewhere about 20, which is not easily interpretable. For these reasons, we decided to discretize the *final_certainty* into three levels (A, B, C). The performance of the certainty levels is discussed later in section 5.3.

- **Level A** contains all the *cz_AoA* with two or more answers and also *cz_AoA* with one answer combined with the *du_freq* estimator. Altogether, 1311 words are of the certainty level A.
- **Level B** contains all words with *du_freq* or *cz_AoA1*, which also do not belong to the certainty level A. It mainly consists of a

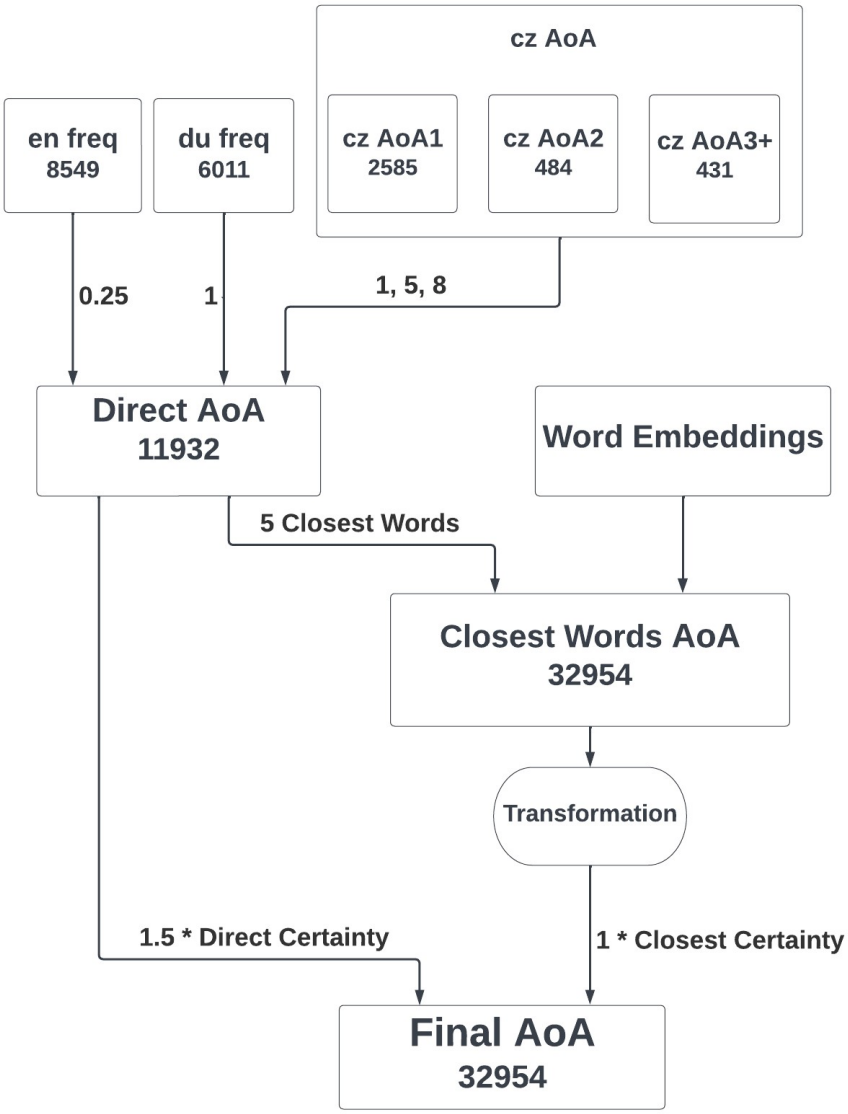

Figure 4.4: Final Estimation. An extension of the diagram for Direct Estimation 4.3.

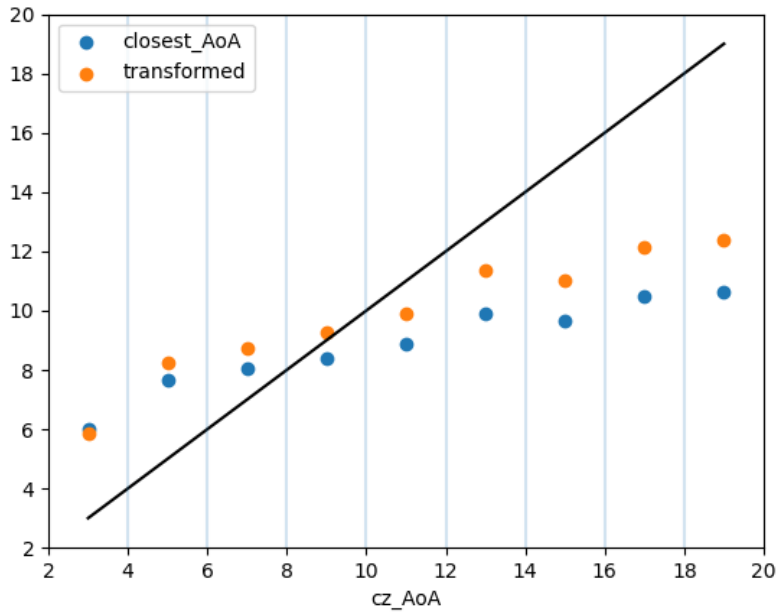

Figure 4.5: Regression to the mean. The dots represent the *closest_AoA* estimation means of all words which have *cz_AoA* in the range of vertical blue lines. The blue dots represent original estimations and the orange ones after the transformation.

combination of *du_freq* and *en_freq* (because these two estimators overlap heavily). The certainty level B is assigned to 7635 words.

- **Level C** includes all other estimated words (24008). An overwhelming majority of them are words without *direct_AoA* at all (21 022). The rest are words with *direct_AoA* created only from estimator *en_freq*. These are words for which we have English AoA but not Dutch AoA (or any other predictor), and there are only about 2986 of them (see the Venn diagram 4.2).

4.3 Regression to the Mean

When averaging AoAs of close words, regression toward the mean occurs. This is caused by the fact that for a word with high AoA ("beret"), its close words will likely include very easy ones ("hat"), which will drag the word's estimation towards the average. This can also happen in reverse (for a word with low AoA, there is a close word with high AoA), but it is much less common. The regression is visible in Figure 4.5. In an ideal case (without regression to the mean), the dots would lie on the black diagonal. The difference between orange and blue dots shows how the transformation affected the data and demonstrates that regression to the mean occurs more in words with higher AoA. To find the appropriate transformation, we used the following steps:

- Use only words with *cz_AoA*. The estimator *cz_AoA* serves as ground truth for transforming the estimations.
- Create a sample of words. From every age group with an interval of 2 years starting at 2 (2-4, 4-6, 6-8, etc.), select 50 words (select all if the group has less than 50 words). This is done because the majority of words are between 7 and 11 years, and using all of them would again drag the estimators towards the mean.
- Find parameters for linear[3] transformation (by grid search or linear regression). We used grid search with mean absolute error as loss function because it penalizes outliers less than mean squared error and thus reduces regression to the mean.

The resulting transformation is $y = 1.4x - 2.5$. Figure 4.6 shows how this transformation affects all the words. We visibly eliminated most regression toward the mean but paid for it by increasing the error overall.

A reader might notice that sampled testing data set was also used when regressing to find coefficients for *en_freq* and *du_freq* estimators. This was also done in order to prevent regression to the mean. If we had used non-sampled training data set, the resulting estimator

3. General logistic curve was also tried, but grid search on the parameters produced a linear curve.

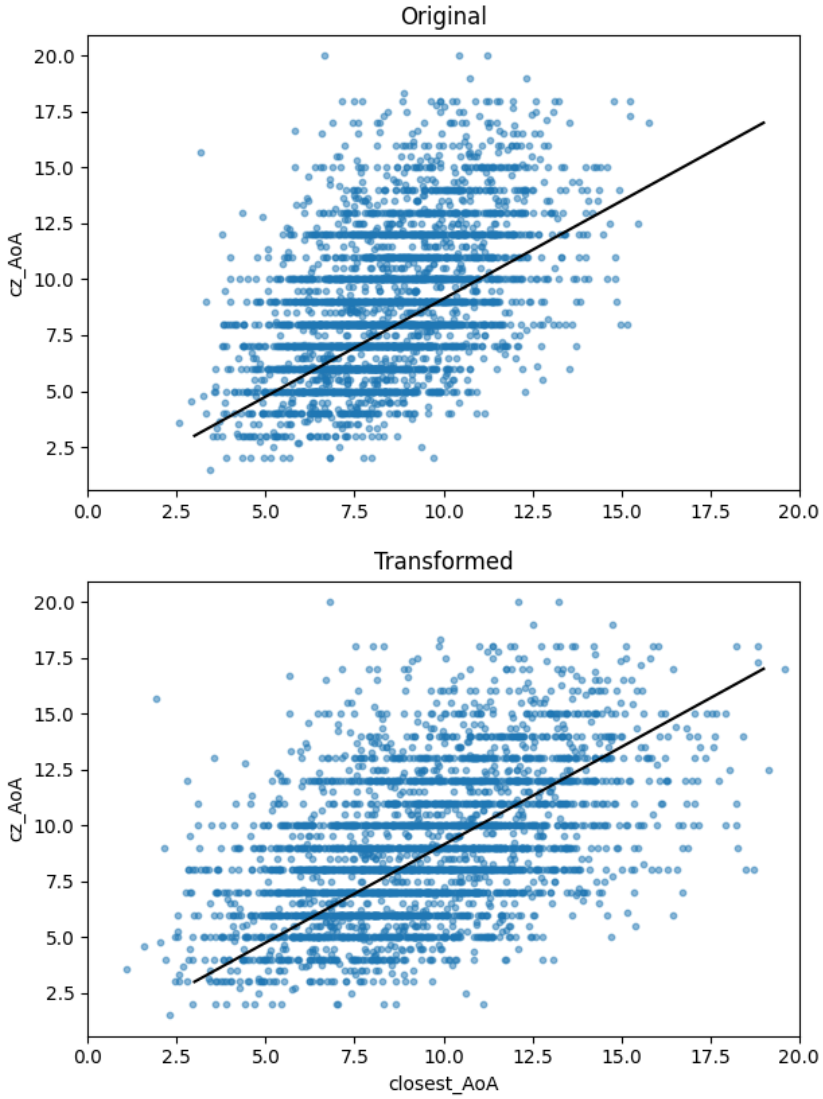

Figure 4.6: Transformation of the closest_AoA estimator.

would strongly favor age groups near nine years (the mean of the distribution). That is because the AoA distribution is not uniform but peaking around nine years; see the distribution of collected data in Figure 3.3.

5 Results

The best possible way to evaluate the results would be to compare them with objective AoA collected for a small subset of words (a few hundred). Unfortunately, such a list is not available for the Czech language, and it was beyond the scope of this thesis to conduct another experiment solely for the purpose of evaluation.

Therefore we decided to evaluate the results in two parts. In the first part, we assume that *cz_AoA* (collected in the experiment) correctly reflects AoA, and we use it as ground truth for every evaluation. This part serves to relatively compare estimators (section 5.2) and to assess the quality of the final prediction and the certainty levels (section 5.3). In the second part, we verify our assumption (that CZECH data are correct) by an evaluation on objective AoA (sections 5.4 and 5.5).

5.1 Evaluation by Czech AoA

The following sections (Comparison of Estimators 5.2, Final Estimator Quality 5.3) are both using *cz_AoA* as an evaluation metric. This section explains how we use Czech answers both as an estimator and evaluator. The evaluation metrics will be Pearson correlation coefficient (Pearson r), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error).

Because the evaluation is done on *cz_AoA*, this part regards only the 3 500 words for which we were able to collect it. Before creating the final estimates, we excluded Czech answers for every one of these words and, from them, created *test_AoA*, which is then used as ground truth. This process is depicted in Table 5.1. Note that in the *test_AoA* column there always has to be at least one answer so that there is *test_AoA* to use for evaluation (if *cz_AoA* is empty, other estimators will be used to create the final estimate).

The splitting of answers is done as follows. 2 586 words with *cz_AoA* have only one answer, so they are all used for the creation of *test_AoA* (for these words, *cz_AoA* will not be used in creating *final_AoA*). For the remaining 914 words with multiple answers, one answer is chosen to be in *test_AoA*; the rest is with a probability of 10% part of *cz_AoA* and with a probability of 90% in the *test_AoA*. The 10% was chosen

Table 5.1: Splitting answers from the experiment for evaluation by *cz_AoA*. Columns cz_AoA and test_AoA contain the answers (left side of the arrow) and the computed mean from them (right side of the arrow).

Slovo	All answers	cz_AoA	test_AoA
betonový	[4, 10, 6, 6]	[10] → 10	[4, 6, 6] → 5.3
včela	[4, 3, 4]	[]	[4, 3, 4] → 3.3
enzym	[15, 14, 16]	[15, 16] → 15.5	[14] → 14
posvítit	[8]	[]	[8] → 8
vepsaný	[8, 9, 14]	[8] → 8	[9, 14] → 11.5

because we have collected *cz_AoA* for 3 500 words, which is approximately 10% of the WORD SET. The disadvantage of this approach is that the *test_AoA* evaluator is not very precise because its estimations mostly consist of only one answer from the experiment, which carries a significant amount of noise. Naturally, this answer splitting is only done for evaluation purposes, and all the data will be used to create the final data set.

5.2 Comparison of Estimators

The goal of this evaluation part is to compare the relative quality of estimators. It is a crucial part of determining the estimator weights and regression coefficients in section AoA Estimators 4.2.

To compare the estimators, we selected only the 487 words for which we are able to use all three of the estimators used in direct estimation (*en_freq*, *du_freq*, *cz_AoA*). These 487 are the intersection of ENGLISH, DUTCH and CZECH (see Venn diagram in Figure 4.2).

The estimators *cz_AoA2* and *cz_AoA3+* are available only for a subset of words, and their reported results (RMSE, correlation) are from these subsets. To maximally use the information available, estimator *cz_AoA1* consists not only from *cz_AoA* with one answer but also

Table 5.2: Quality of estimators. The column Weight refers to the weight in weighted average when combining estimators.

Estimator	RMSE	Pearson r	Size	Weight
reg(*length*)	3.81	0.34	487	
reg(*log_freq*)	3.52	0.47	487	
reg(*en_AoA*)	3.08	0.64	487	
reg(*du_AoA*)	3.01	0.69	487	
en_freq	3.04	0.65	487	0.25
du_freq	2.88	0.7	487	1
cz_AoA1	2.92	0.73	487	1
cz_AoA2	2.29	0.82	95	5
cz_AoA3+	1.73	0.9	51	8

from *cz_AoA* with more answers from which one answer is randomly selected. [1]

The results are shown in Table 5.2. According to our expectations, the *reg*(*length*) scored the worst and *cz_AoA3+* the best. The estimator *du_freq* seems to be comparable in quality to *cz_AoA1*, and we can also see a considerable drop in RMSE between *cz_AoA1* (2.92) and *cz_AoA2* (2.29), which agrees with the analysis of RMSE in Chapter 3. Although the evaluation is on only 487 words, the correlations to *cz_AoA* are the same when testing with the maximum possible words for the estimators.

5.3 Final Estimator Quality

The following evaluation is done on the 3500 test estimations derived from *cz_AoA*. The results are displayed in Table 5.3. Evaluated are not only all estimations but also the certainty levels. As expected, certainty level A shows the best results and certainty level C the worst. The quality difference between levels is more clearly visible in Figure 5.1. The scatter plots also show that the estimations do not appear to have

1. Similar to *cz_AoA2*, it consists of *cz_AoA* with two answers or *cz_AoA3+* from which are two answers randomly selected.

Table 5.3: Quality of *final_AoA*.

Certainty	Pearson r	RMSE	MAE	Size
All	0.56	2.99	2.35	3500
A	0.89	1.90	1.31	62
B	0.73	2.68	2.1	644
C	0.49	3.07	2.45	2794

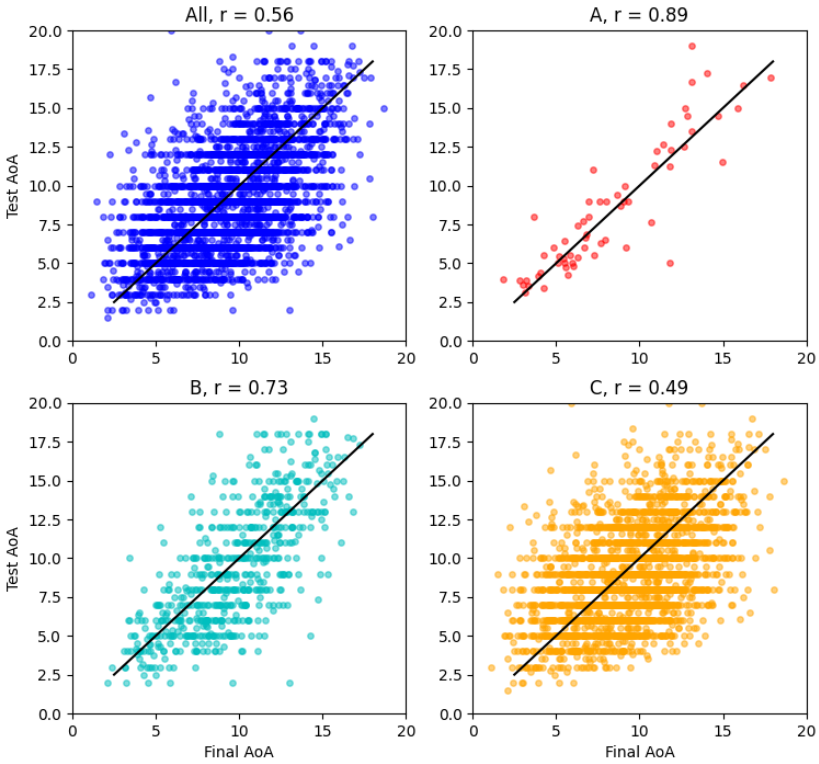

Figure 5.1: Scatter plots of certainty levels. The first scatter plot shows all the tested words and the other 3 show only words from a certainty level (A, B, C). The value of Pearson correlation is included in every plot.

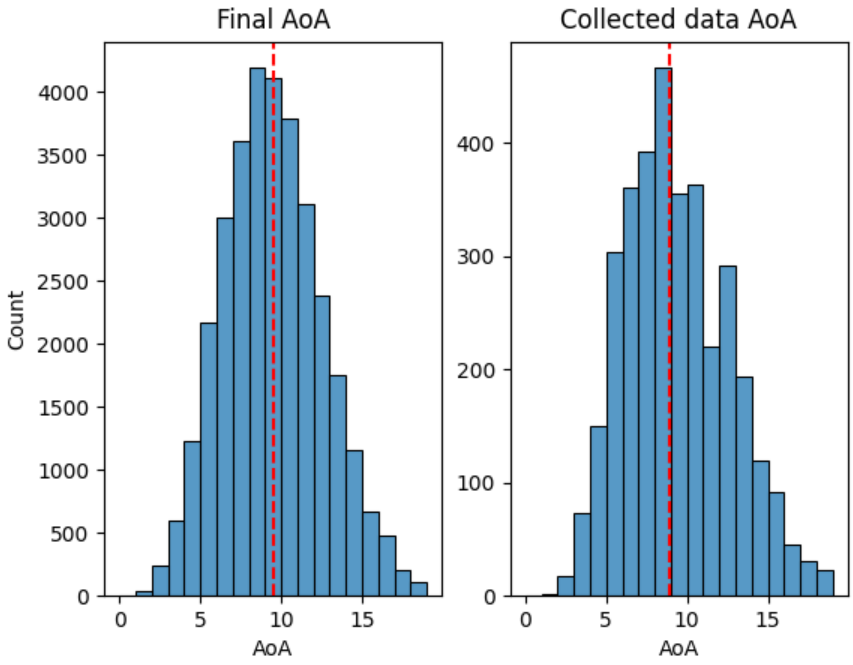

Figure 5.2: Distributions. The figure shows the difference in distributions between the AoA of collected data (which serves as ground truth) and the *final_AoA*.

visible regression to the mean or other unwanted phenomenons except dispersion which decreases with better quality levels.

5.3.1 Distribution

The distribution of predicted data is normally shaped with a mean of 9.4 and slightly skewed to the right. The Figure 5.2 is a comparison of the *final_AoA* and the collected data distributions. The ideal result would be if the two images were identical, which would mean that our estimator mirrors the age distribution of the collected data. One can see that the distribution on the left is slightly steeper. That is caused by the occurring regression to the mean, which we tried to eliminate by transformation.

5.4 Evaluation by Calibration Table

For the Czech data collection purposes, we collected semi-objective AoA for 57 Czech words. The words' AoA were initially created by us and then edited based on the knowledge of a small number of children. The evaluation on these words is possible only because they were excluded from the experiment. The words are among the more common ones; therefore, for the majority of them, English and Dutch AoA were available. That causes only a small part of them to be of the certainty level C. There are also no words with level A because level A requires *cz_AoA* (collected in the experiment from which these words were excluded). The results are visible in the Figure 5.3. This evaluation serves mainly as a confirmation that we are able to predict objective evaluations (because up to this point, we were evaluating only by subjective AoA collected in the experiment).

5.5 Experiment with Children

The last experiment to confirm the usefulness of the created estimations was done by quizzing 4 children on 112 words selected[2] from the WORD SET. It is a variant of evaluation by objective AoA where data collection is not as time-consuming.

The children were asked to explain the meaning of every word. The age-of-acquisition of a word is defined as that 50% of children of that age have the word acquired, it implies that a child should on average have acquired 50% of words with the AoA of the same age. We can estimate accuracy of our results by how precisely we are able to guess a child's age based on his/her answers.

Figure shows the experiment with four children. The results for all four children, visible in Figure 5.4, were relatively accurate. Based on which words the children knew, one could guess their age with precision to half a year (approximately). For example, the Child C is 10.5 years old, which means that ideally, he/she should have acquired more than 50% of words with AoA 10 and less than 50% of words with AoA 11, which is exactly what the figure shows. One can see that the results for the other three children are similarly successful.

2. The selection was made by stratified sampling to include all age groups.

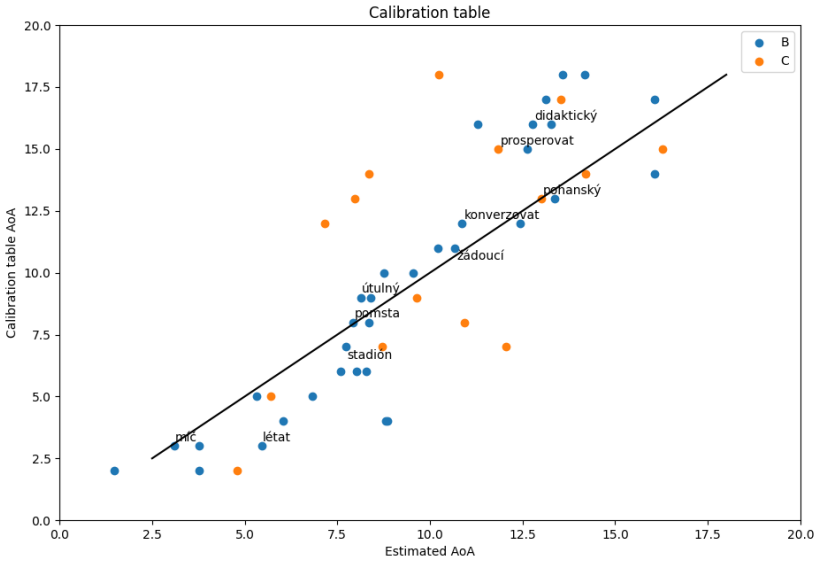

Figure 5.3: Calibration table evaluation. Compared are the final estimations and the AoA from Calibration table. The blue dots represent words with certainty level B and the orange dots words with certainty level C.

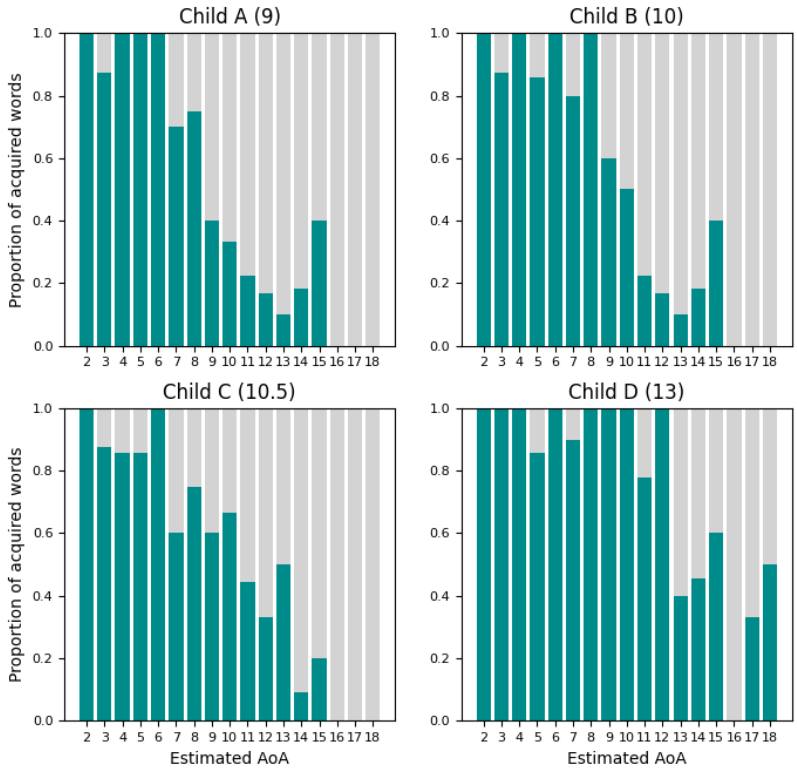

Figure 5.4: Answers from children. On the x-axis is AoA of asked words, and on the y-axis is a proportion of of words the child has acquired.

6 Discussion

The resulting data set of 32 954 words was created by combining multiple estimators from different sources. Because the available sources differed word by word, we created 3 certainty levels which indicate how confident we are about the estimate. Due to a lack of sources, the majority of estimates (24 008 of 32 954) are of level C (the worst one), which combines AoA estimations of its 5 closest words. The results for certainty levels A and B showed that the estimation accuracy increases significantly with better/more sources. From our analyses, we deduced that words of certainty level B are quality-wise similar to an estimate from a native speaker (level A is slightly better, and level C noticeably worse).

The motivation for this thesis was to use the data set of estimations as a tool for authors of text comprehension exercises. Based on the positive results of experiments using the objective AoA (Calibration table experiment, Experiment with children), we are confident to say that for this purpose, the results are comfortably usable. We imagine the following use: from the text, find words with (significantly) higher AoA than the age of the expected reader. Present these words to the text author together with the AoA estimations and the accuracy of the estimation. We believe that for the author, the relevancy of found words should be (at least partly) based on their estimation certainty.

Another possible use is for teaching Czech to foreigners. The data set can be used to estimate difficulty and create a list of words adequate to the level of the student. We would again advise using the estimates more as a suggestion, in this case not only because of their (in)accuracy but also because AoA and the difficulty of words for foreign speakers often differ (e.g. the word yummy is of low AoA but is not a basic word when learning English).

We also think that there is potential in using the words of certainty level A in studies of other psycholinguistic variables which closely correlate with AoA (e.g. familiarity, imageability) and, of course, in any follow-up studies of Czech AoA.

In summary, all estimations can be used as a preprocessing tool (presorting, filtering, selecting), and the estimations of level A are usable as a psycholinguistic metric.

6.1 Czech AoA Findings

Apart from the main goal of data set creation, this thesis also serves to broaden the knowledge about Czech AoA. We found that the length and frequency of the word correlate with AoA by rates of 0.34 and -0.47, respectively. These results are in agreement with a study from Łuniewska et al. [8], who reported that Czech AoA correlates by 0.33 to length and -0.45 to frequency. Their results also included findings on the correlation among different languages. Their reported correlation between Czech and English was 0.77 and Czech with Dutch by 0.85. We found smaller correlations of 0.65 and 0.70, respectively, but the difference might be caused by the fact that we did not use objective evaluations as a scoring metric. We also found that subjective evaluations are surprisingly stable. The average of 3 answers correlates with our test AoA by a factor greater than 0.9.

6.2 Limitations and Future Work

There are clearly ways how to improve on these results. The greatest drawback of the thesis is the insufficient amount of collected data (3 500 words). It allowed us to create a rough estimation and analyze different techniques, but it was not enough to cover all words in the Czech language even with the use of other sources.

The quality of collected data is also an issue; the majority of estimations (2 586) consisted of only one answer, and there was close to no control over how the experiment was distributed and who participated in it. We suppose that the majority of answers came from students and other people close to the academic sphere, and that could have influenced the results. To achieve more reliable results, a reward for participating should be administered (as was for other similar studies), and information about participants (age, education) should be collected. From our analysis, we know that a single answer is of similar quality as translations from Dutch, and the average of 2 or 3 answers improves the accuracy greatly. These results imply that a promising direction for a follow-up study is to gather 2-3 answers for every word.

One other possible direction of follow-up research is the utilization of the Czech acquisition corpora. The most promising corpus we found was SKRIPT 2012 [25], which contains transcriptions of Czech students' essays from years 10 to 19 approximately and consists of 708 668 lemmas. It could be a major source of objective AoA and, therefore, very valuable to both estimation and evaluation.

Bibliography

1. BALYAN, Renu; MCCARTHY, Kathryn; MCNAMARA, Danielle S. Applying Natural Language Processing and Hierarchical Machine Learning Approaches to Text Difficulty Classification. *International Journal of Artificial Intelligence in Education*. 2020, pp. 1–34. Available from DOI: 10.1007/s40593-020-00201-7.
2. JUHASZ, Barbara J. Age-of-acquisition effects in word and picture identification. *Psychological bulletin*. 2005, vol. 131, no. 5, p. 684. Available from DOI: 10.1037/0033-2909.131.5.684.
3. KUPERMAN, Victor; STADTHAGEN-GONZALEZ, Hans; BRYSBAERT, Marc. Erratum to: Age-of-acquisition ratings for 30,000 English words. *Behavior research methods*. 2012, vol. 44. Available from DOI: 10.3758/s13428-012-0210-4.
4. BRYSBAERT, Marc; STEVENS, M.; DE DEYNE, Simon; VOORSPOELS, Wouter; STORMS, Gert. Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*. 2014, vol. 150, pp. 80–84. ISSN 0001-6918. Available from DOI: 10.1016/j.actpsy.2014.04.010.
5. IMBIR, Kamil K. Affective norms for 4900 Polish words reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Frontiers in psychology*. 2016, vol. 7, p. 1081. Available from DOI: 10.3389/fpsyg.2016.01081.
6. MORRISON, Catriona; CHAPPELL, Tameron; ELLIS, Andrew. Age of Acquisition Norms for a Large Set of Object Names and Their Relation to Adult Estimates and Other Variables. *The Quarterly Journal of Experimental Psychology*. 1997, vol. 50A, pp. 528–559. Available from DOI: 10.1080/027249897392017.
7. BIEMILLER, Andrew; ROSENSTEIN, Mark; SPARKS, Randall; LANDAUER, Thomas K.; FOLTZ, Peter W. Models of Vocabulary Acquisition: Direct Tests and Text-Derived Simulations of Vocabulary Growth. *Scientific Studies of Reading*. 2014, vol. 18, no. 2, pp. 130–154. Available from DOI: 10.1080/10888438.2013.821992.

8. ŁUNIEWSKA, Magdalena; HAMAN, Ewa; ARMON-LOTEM, Sharon; ETENKOWSKI, Bartłomiej; SOUTHWOOD, Frenette; ANĐELKOVIĆ, Darinka; BLOM, Elma; BOERMA, Tessel; CHIAT, Shula; ABREU, Pascale Engel de, et al. Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior research methods*. 2016, vol. 48, no. 3, pp. 1154–1177. Available from DOI: 10.3758/s13428-015-0636-6.

9. BRYSBAERT, Marc; BIEMILLER, Andrew. Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior research methods*. 2016, vol. 49. Available from DOI: 10.3758/s13428-016-0811-4.

10. RUSSO, Irene. Guessing the age of acquisition of italian lemmas through linear regression. In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. 2020, pp. 43–48. Available from DOI: 10.18653/v1/2020.cmcl-1.5.

11. ŁUNIEWSKA, Magdalena; WODNIECKA, Zofia; MILLER, Carol A; SMOLÍK, Filip; BUTCHER, Morna; CHONDROGIANNI, Vasiliki; HREICH, Edith Kouba; MESSARRA, Camille; A. RAZAK, Rogayah; TREFFERS-DALLER, Jeanine, et al. Age of acquisition of 299 words in seven languages: American English, Czech, Gaelic, Lebanese Arabic, Malay, Persian and Western Armenian. *PloS one*. 2019, vol. 14, no. 8, e0220611. Available from DOI: 10.1371/journal.pone.0220611.

12. HANSEN, Pernille. What makes a word easy to acquire? The effects of word class, frequency, imageability and phonological neighbourhood density on lexical development. *First Language*. 2017, vol. 37, no. 2, pp. 205–225. Available from DOI: 10.1177/0142723716679956.

13. PAETZOLD, Gustavo; SPECIA, Lucia. Inferring Psycholinguistic Properties of Words. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016, pp. 435–440. Available from DOI: 10.18653/v1/N16-1050.

14. SOARES, Ana Paula; COSTA, Ana Santos; MACHADO, João; COMESAÑA, Montserrat; OLIVEIRA, Helena M. The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. *Behavior Research Methods*. 2017, vol. 49, no. 3, pp. 1065–1081. Available from DOI: 10.3758/s13428-016-0767-4.

15. CAMEIRAO, Manuela L; VICENTE, Selene G. Age-of-acquisition norms for a set of 1,749 Portuguese words. *Behavior research methods*. 2010, vol. 42, no. 2, pp. 474–480. Available from DOI: 10.3758/BRM.42.2.474.

16. MIKOLOV, Tomas; GRAVE, E.; BOJANOWSKI, P.; PUHRSCH, Christian; JOULIN, Armand. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*. 2017. Available from DOI: 10.48550/arXiv.1712.09405.

17. MANDERA, Paweł; KEULEERS, Emmanuel; BRYSBAERT, Marc. How useful are corpus-based methods for extrapolating psycholinguistic variables? *Quarterly journal of experimental psychology (2006)*. 2015, vol. 68, pp. 1–20. Available from DOI: 10.1080/17470218.2014.988735.

18. BOTARLEANU, Robert-Mihai; DASCALU, Mihai; WATANABE, Micah; MCNAMARA, Danielle S; CROSSLEY, Scott Andrew. Multilingual Age of Exposure. In: *International Conference on Artificial Intelligence in Education*. 2021, pp. 77–87. Available from DOI: 10.1007/978-3-030-78292-4_7.

19. KŘEN, Michal; CVRČEK, Václav; ČAPKA, Tomáš; ČERMÁKOVÁ, Anna; HNÁTKOVÁ, Milena; CHLUMSKÁ, Lucie; JELÍNEK, T.; KOVÁŘÍKOVÁ, Dominika; PETKEVIC, Vladimir; PROCHÁZKA, Pavel, et al. SYN2015: Representative corpus of contemporary written Czech. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 2522–2528. Available also from: https://aclanthology.org/L16-1400.

20. BIČAN, Aleš. *Fonologický lexikální korpus češtiny a slabičná struktura českého Slova* [https://ujc.avcr.cz/phword/]. [N.d.]. Accessed: 2022-03-06.

21. *Age-of-acquisition (AoA) norms for over 50 thousand English words* [http://crr.ugent.be/archives/806]. [N.d.]. Accessed: 2022-03-14.

22. BIRCHENOUGH, Julia MH; DAVIES, Robert; CONNELLY, Vincent. Rated age-of-acquisition norms for over 3,200 German words. *Behavior research methods*. 2017, vol. 49, no. 2, pp. 484–501. Available from DOI: 10.3758/s13428-016-0718-0.

23. GRAVE, Edouard; BOJANOWSKI, P.; GUPTA, Prakhar; JOULIN, Armand; MIKOLOV, Tomas. Learning Word Vectors for 157 Languages. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018. Available from DOI: 10.48550/arXiv.1802.06893.

24. *Kategorie: Vulgární Výrazy/čeština – wikislovník*. [N.d.]. Available also from: https://cs.wiktionary.org/wiki/Kategorie:Vulg%C3%A1rn%C3%AD%20v%C3%BDrazy/%C4%8De%C5%A1tina.

25. ŠEBESTA, Karel; GOLÁŇOVÁ, Hana; JELÍNEK, T.; JELÍNKOVÁ, Blanka; KŘEN, Michal; LETAFKOVÁ, Jana; PROCHÁZKA, Pavel; SKOUMALOVÁ, Hana. *SKRIPT2012: akviziční korpus psané češtiny–přepisy písemných prací žáků základních a středních škol v ČR*. Ústav Českého národního korpusu, 2013. Available also from: http://www.korpus.cz.

A Electronic Attachments

A.1 AoA Data Set

The primary result of the thesis. The file `results.csv` contains 32 954 items with Czech words, estimated AoA, and the estimation certainty.

A.2 Experiment Source Code + Log

The zip file `experiment.zip` contains a source code for the experiment discussed in Chapter 3 and a log file containing all the collected data.

A.3 Filtered Words and Logs

The file `filtered.zip` containing information that we decided to exclude from our analyses. The file `filtered_logs.csv` contains a list of logs from the experiment which we considered mistakes from participants. The file `filtered_words.csv` contains a list of words (grouped logs by words) that had too high a deviation between individual answers.

A.4 Word Set Creation Source Code

Attachment `word_set_creation.zip` contains scripts and data used to create the word set. There is also a diagram `word_set_diagram.png` which shows the overview of how is the data combined.

A.5 Analyses

Attachment `analyses.zip` contains the source code of creating the data set and relevant analyses in this thesis.

B Calibration Table

Table B.1: Calibration table

věk	podstatné	přídavné	sloveso
2	maminka	mokrý	spát
3	míč	modrý	létat
4	oslava	deštivý	zápasit
5	obdélník	skvělý	vyplnit
6	vězeň	statečný	zklamat
7	stadión	opadavý	předpovědět
8	pomsta	mělký	otupit
9	buňka	útulný	navyknout
10	vdova	tradiční	vzkřísit
11	tlumočník	žádoucí	tápat
12	nárt	blažený	konverzovat
13	pokrytectví	pohanský	znehodnotit
14	ručitel	bederní	zapudit
15	kýč	jízlivý	prosperovat
16	pranýř	didaktický	vykořisťovat
17	patologie	zhrzený	stylizovat
18	hegemonie	emeritní	zpronevěřit