ABSTRACT

In this study Age of Acquisition (AoA) of words is predicted using different word frequency distributions and other word attributes, like word length, word concreteness, lexical decision times, word arousal, valence and dominance, prevalence and word frequency. Two analysis models are used: Multiple Regression Analysis (MRA) and Random Forests. Both analyses are performed in both Dutch and English, but not every predictor is used in both languages. Furthermore different word frequency variables are used in both languages. These variables represent the word frequency distribution from different subsets of natural language (i.e., written, spoken, and child-directed). The results show that Random Forests better predict AoA than MRA and that in that case child-directed word frequency variables are better for predicting AoA in both Dutch and English. Furthermore there is a very strong effect of Concreteness and Lexical Decision Time on AoA.

INTRODUCTION

CHILDERNS’ WORD LEARNING PROCESS

Our abilities to speak, write, and read are some of our most used skills every day and they form an integrated whole with language. Language is a dynamic concept and has many functions, it is for example essential for thinking, expressing oneself and communicating with others. Unfortunately knowing a language is not innate, so we need to learn it. This takes many years, because language has many aspects, like vocabulary, pronunciation, grammar, and sentence structure. Actually you never stop improving and enlarging your knowledge of your native language.

Learning words is a very important aspect of the language acquisition process. This already starts at an early age of 12 weeks, when children start producing sounds in their native language, referred to as babbling (Kuhl, 2004). 75% of the children produce their first words before their first birthday (Schneider, Yurovsky, & Frank, 2015). Humans continue learning new words during their whole life, but during childhood their vocabulary grows with the highest rate.

Word learning is an individual process and children do not learn every word at exactly the same age, but various researchers showed that there exists a rough learning order of words (Brysbaert, Stevens, De Deyne, Voorde, & Storms, 2014a; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). Schneider et al. (2015) concluded that children between 8 and 30 months old start with learning roughly the same words. Furthermore Luniewska et al. (2016) found that this learning order is also cross-linguistic, this means that words with similar meanings among languages are learned at roughly the same age.
Experiments show that children are able to learn word meanings from single instances (Markson & Bloom, 1997; Spiegel & Halberda, 2011), but also from cross-situational mappings (Smith & Yu, 2008). In a simple example of a cross-situational mapping the learner sees two objects and hears two unknown words, for example BAT and BALL. It is unclear which word belongs to which object. In the second trial the learner sees one new object and one object from the first trial and again two words are called, for example BALL and DOG. Now the word BALL can be associated with the object that the learner saw both in trial 1 and trial 2 (Smith & Yu, 2008). It is not known whether learning from single instances or from cross-situational mappings is used most when learning words and whether most words are acquired using many instances of the words or just one or a few instances are enough. It is certain that a child needs to hear the word referring to a certain concept at least once, because it never happens that a child starts speaking Chinese while in its environment only English is spoken. A measure that represents how often a word occurs in a language is word frequency. Zipf (1949) discovered that the frequency of a word in a given corpus of natural language is inversely proportional to its rank in the frequency table. Many words are rare and learning them requires extended exposure to language (Landauer & Dumais, 1997). Every exposure to a word represents a learning opportunity of this word, so word frequency is an interesting variable to take into account for predicting the word learning process.

Besides word frequency there are many other input-related word attributes that may influence the word learning order, like word length, initial phoneme, co-occurrence with other words, lexical density, and orthographic and phonological distance to neighbors. A neighbor is defined as a word that differs from the target word in just one letter position (Schepens, Dijkstra, Grootjen, & van Heuven, 2013). Also conceptual attributes like semantic similarity, familiarity, babiness (the extend in which a word is relevant for babies), concreteness, and emotional value (how pleasant, dominant or arousing a word is) can be of influence.

Adamson et al. (2016) showed that parents adjust their speech to suit their child’s level of word learning. This means that their speech has a different word frequency distribution compared to the natural language word frequency distribution. Therefore the first research question is: What word frequency distribution from which subset of natural language (i.e., written, spoken, or child-directed) is the best predictor of the word learning order? Furthermore I am also interested in the individual relationship between other word attributes and AoA, so my second question is: “To what extent are certain input-related and conceptual factors good predictors of the age at which words are learned?” Both questions will be answered twice, once with evidence based on English data and once with evidence based on Dutch data.

HOW TO MEASURE THE WORD LEARNING PROCESS?
To measure the word learning process we need information about the average age at which children learn words, the age of acquisition (AoA). This is a difficult variable to measure. Until now researchers have obtained data about AoA in two ways, a so-called quasi-objective way and a subjective way. Of course acquiring data about AoA in an objective way is most desirable, but unfortunately this is also very difficult, because it is hard and time-consuming for parents and other observers to determine the exact moment in time when a child learns a word.

The MacArthur-Bates Communicative Development Inventory (MB-CDI) (Frank, Braginsky, Yurovsky, & Marchman, 2016) is a data collection from which quasi-objective AoA scores can
be determined (Łuniewska, et al., 2016). To collect this data researchers give parents a list containing 299 words and ask them to mark the words on the list that their child understands and/or produces. Unfortunately this approach is biased when used for research about the word learning process, because only information about the 299 preselected words is obtained. Furthermore this data collection still depends on estimations made by parents, but it is proved that the MB-CDI is a highly reliable database (Heilmann, Ellis Weismer, Evans, & Hollar, 2005).

In the subjective approach adult participants were presented a word and they had to decide at which age they thought children should learn that word. Different studies have obtained data about AoA using this method (Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014a; Łuniewska, et al., 2016; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012).

Many researchers studied the relationship between objective and subjective AoA. Gilhooly et al. (1980) collected quasi-subjective AoA ratings by taking a vocabulary test with subjects from 5 till 21 years old. The same words were also rated by adults and after comparing both they found a high correlation between the ratings of the two groups (r = 0.84). Morrison et al. (1997) also found a high correlation between the quasi-objective AoA ratings and adult ratings (r = 0.76) (Morrison, Chappell, & Ellis, 1997). Both concluded that they prefer objective AoA ratings when available, but subjective adult ratings provide also a reliable and valid measure of real word learning age and is definitely a good alternative.

This study uses two collections of AoA ratings, one in Dutch and one in English. These ratings are all obtained in a subjective way by asking adults at which age they thought they had learned the word. Luniewska et al. (2016) compared this question to the more general question: “When do you think children learn this word?” and found that participants reporting their own experience provided significantly higher AoA ratings then those reporting when children in general learn the word. Kuperman et al. (2012) collected the first collection of AoA ratings by using an online survey. This collection contains AoA ratings for 31,124 English words. They validated their data by comparing the ratings to three earlier studies with smaller amounts of AoA ratings that were obtained under controlled circumstances.

Brysbaert et al. (2014) collected AoA ratings for around 25,888 Dutch words. To obtain this data they divided these words into five lists of 6,000 words each and every list was rated by 15 students from Ghent university. Moors et al. (2013) proved that this amount of raters is enough to deliver suitable AoA ratings.

PREDICTORS OF AGE OF ACQUISITION

Various possible predictors of AoA were discussed in the introduction and just a few of them will be included in the analysis to discover their effect on AoA. In this section these selected variables will be described in more detail, together with the English and Dutch sources where the values for the variables were achieved from and the expected effect from the variable on AoA.

WORD LENGTH

Lewis et al. (2016) found that participants mapped longer words to more complex objects in language comprehension and language production tasks. (Lewis & Frank, 2016). Furthermore, longer words are also more difficult to remember and produce, in the first place because word complexity correlates with word length (Lewis & Frank, 2016), but also because there are more characters and more phonemes. Because more complex words have a relatively higher AoA
than less complex words, this makes it interesting to study the relationship between word length and AoA. This variable can be measured in number of characters, number of syllables or number of phonemes. The syllable counts in this study are achieved from The Dutch Lexicon Project 2 (Brysbaert, Stevens, Mandera, & Keulers, 2016) and The British Lexicon Project (Keulers, Lacey, Rastle, & Brysbaert, 2012), containing syllable data for respectively 29,956 and 28,550 words. I expect that a positive relationship exists between word length and AoA. This means that the more characters or syllables a word has, the older the child is when it is learned.

LEXICAL DECISION TIMES
The AoA effect states that words acquired earlier in life are processed faster and more accurate than words acquired later in life (Juhasz, 2005). The processing time of a word is often measured using a lexical decision task. In such a task a participant is exposed to a stimulus and he/she has to respond on this stimulus. For example: a written word appears on the screen and the participant has to press the left button if the word exists in his/her native language and the right button if the word does not exist. Kuperman et al. (2012) reported a high correlation between AoA and lexical decision tasks. The Dutch lexical decision times were achieved from the Dutch Lexicon Project 2 (Brysbaert, Stevens, Mandera, & Keulers, 2016). This project contains lexical decision times for 59,616 words. The English lexical decision times were achieved from the British Lexicon Project, containing data for 55,867 words (Keulers, Lacey, Rastle, & Brysbaert, 2012). I expect a positive relationship between lexical decision time and AoA: the higher a word’s lexical decision time, the older the child is when the word is learned.

WORD CONCRETENESS
Since Paivio (1971) formulated his dual-coding theory in which he stated that more concrete words are easier to remember than abstract words (Paivio, 1971), researchers start collecting concreteness ratings and the variable concreteness starts being used in psycholinguistic research. Concreteness ratings, also referred to as imageability ratings (Connell & Lynott, 2012), are collected by asking participants to rate on a scale how abstract or concrete they think a certain word is. In this study I will use the largest collections of concreteness ratings available at the moment. The English ratings consists of 39,954 generally known English word lemmas. These words were rated on a 5-point scale, where 1 means abstract (language based) and 5 means concrete (experience based) (Brysbaert et al., 2014b). The Dutch database contains 30,070 Dutch words and the ratings were achieved on the same scale as in the English version (Brysbaert et al., 2014a). The meaning of a concrete word is most of the time very easy. Furthermore it is possible to point to such words in most cases. That is why I expect that there exists a negative relationship between concreteness and AoA, what means that concrete words (with a higher rating) are learned earlier than abstract words (with a lower rating).

WORD FREQUENCY
Since Howes and Solomon (1951) discovered that word frequency influences cognitive processing, many studies have been gathering word frequency data based on different corpora to study the word frequency effect. This effect is that high frequency words are processed faster and more efficiently than low-frequency words (Ferrand, et al., 2011; New, Brysbaert, Veronis, & Pallier, 2007). Furthermore high-frequency words are easier to recall, but more difficult to recognize in episodic memory tasks (Glanzer & Bowles, 1976). Earlier research has revealed that word frequency seems to be one of the strongest predictors of AoA (Braginsky, Yurovsky, Marchman, & Frank, 2016; Schneider, Yurovsky, & Frank, 2015).
When writing an informal letter, different language is used than when writing an academic article. The same applies in spoken language, when adults are talking to children they adjust their speech to the child’s level of word learning (Adamson, Bakeman, & Brandon, 2016). This means that there are different kinds of language depending on the form (i.e., spoken or written language) and the context in which it is used. Every kind of language has its own word frequency distribution. The most objective way to achieve the word frequency distribution is to count the number of times a word occurs in a certain corpus containing written texts or transcribed audio fragments from the desired kind of language. In this study I will consider four English and four Dutch word frequency distributions based on different kinds of languages to discover which distribution best predicts AoA.

English word frequency variables
The first English word frequency measure is extracted from the CHILDES Parental Corpus and consists of 24,156 unique utterances of parents, caregivers and experimenters extracted from the original CHILDES corpus. Although the utterances are not all child-directed, they form a representative sample of spoken words where young children are typically exposed to (MacWhinney, 2000; Li & Shirai, 2000).

The next variable is based on the Google Corpus. This word frequency distribution contains 333,333 words and their frequencies obtained from public webpages (Google Research, 2006). Because of the fact that this frequency distribution is based on a very large corpus, this variable seems a good approach of estimating the true word frequency distribution in natural language.

The third English variable is based on the SUBTLEX-US Corpus. This corpus contains 60,384 unique words in American English. These words and their frequencies were obtained from the subtitles of 8,388 US films and television series that were released between 1900 and 2007 (Brysbaert & New, 2009). New et al. (2007) proved that a corpus based on subtitles is the best approximation of word frequencies in spoken human interactions (New, Brysbaert, Veronis, & Pallier, 2007).

The fourth and last English word frequency variable is the summed frequency. This variable sums for every word the scaled frequencies of the CHILDES, SUBTLEX-US and Google word frequency.

Dutch word frequency variables
In the Dutch part of this study are only two corpora used: BasiLex and SUBTLEX-NL. The BasiLex corpus contains 19,450 unique words. This corpus consists of written texts offered to children in kindergarten, the different classes of elementary school and the first two classes of high school (Tellings, Hulsbosch, Vermeer, & Van Den Bosch, 2014). SUBTLEX-NL is a corpus containing Dutch subtitles from 8,443 films and television series. The majority (5,966) of these subtitles were translated from English films and television series (Keuleers, Brysbaert, & New, 2010).

Based on these two corpora I created four word frequency variables. The first variable is called BasiLexYoung and contains the word frequency distribution from texts offered to kindergarten and group 1 of the elementary school. The second variable, BasiLexAll, contains the frequency distribution over all texts in the BasiLex corpus. The third variable contains the word frequency counts for all words in the SUBTLEX-NL corpus and the last variable is the sum of the normalized BasiLexAll and SUBTLEX-NL frequency counts.
Most children start learning words from spoken conversations with and around them (Kuhl, 2014), so that implies that the CHILDES word frequency distribution best predicts the age when words are learned. But on the other hand is the Google word frequency distribution based on a much larger corpus and more data yields a distribution that approaches the real distribution.

The BasiLex corpus contains texts that are specially adapted to children’s level of language comprehension, so I expect that the variables obtained from this corpus are better in predicting the age of acquisition of words than the SUBTLEX-NL or Summed frequency variable. BasiLexAll is based on all texts in the BasiLex corpus and BasiLexYoung just on texts written for children from kindergarten and the first class of elementary school. Because of this I expect that BasiLexAll is a better predictor, since word learning does not stop after first class of elementary school.

Last but not least I expect a negative relationship between AoA and word frequency. This means that the younger a child is when he/she learns a word, the higher the word frequency of that word is.

**WORD AROUSAL, VALENCE AND DOMINANCE**

Three emotional variables will be included in the English part of the analysis. The first variable is arousal, this is defined as the intensity of emotion provoked by a stimulus. The second variable is valence, being the pleasantness of a stimulus. The last emotional variable is dominance, referring to the dominance or power of a word. In this study the valence, arousal and dominance ratings collected by Warriner et al. (2013) will be used. This dataset consists of 13,915 English words. These ratings were obtained on a scale from one (valence=happy, arousal = excited, dominance = strong/dominant) to nine (valence=unhappy, arousal = calm, dominance = weak/submissive) (Warriner, Kuperman, & Brysbaert, 2013). We expect that there exists a positive relationship between these emotional variables and AoA, so the happier, the more exciting or the stronger/more dominant a word is, the earlier it is learned.

In Dutch only a small collection of word valence, arousal and dominance ratings are available, so these variables will not be part of the Dutch analysis in this study.

**PREVALENCE**

Prevalence is defined as the familiarity of a population with a word. It is measured as the percentage of the population that knows the word. There are prevalence ratings available for 54,319 Dutch words (Keuleers, Stevens, Mandera, & Brysbaert, 2015) I expect a negative relationship between prevalence percentages and AoA exists. This means that the higher the percentage of the population knowing the word, the earlier it is learned. Unfortunately, these ratings have not been collected in English yet. So this variable can only be taken into account in the Dutch part of this study.

**METHOD**

Two datasets, one containing all Dutch variables (N=190,275) and one containing all English variables (N=488,089), are created. Both datasets contain a variable for AoA, word length in number of characters, word length in number of syllables, concreteness, lexical decision time, child-directed frequency (respectively CHILDES-frequency and BasiLexYoung-frequency), subtitle-based word frequency (respectively SUBTLEX-US and SUBTLEX-NL) and summed
frequency. Furthermore the English dataset also contains the three emotional variables arousal, valence and dominance and Google word frequency. The Dutch dataset contains prevalence and BasiLexAll-frequency as extra variable.

The second step is to make sure that there are no double words in the dataset. This is done by formatting all words in the same way (with only lowercase letters) and then summing the word frequency data for double words (not for other variables, because for example length does not become double as long only because the word is twice in the dataset).

The next step is to select words from the big datasets for the analyses. Two important aspects to keep in mind during this selection is that the size of the dataset needs to be as big as possible, but the datasets also need to be mutually comparable. To meet the last aspect, all words having one or more missing values are removed from the big (English or Dutch) dataset to make sure that the word has a value for all four frequency measures and every other variable. This results in two cleaned base datasets in English and Dutch that are large enough (respectively N = 4305 and N= 19,219), so both word selection requirements are met.

Next normalize the predictor values in both base datasets, in this way all variables will have scores on the same scale. Because the effect of different word frequency variables on AoA is measured, this cannot be done in just one analysis using one dataset. So the last step is to create four sub-data sets from every base dataset. Every sub-data set contains the same variables, except for word frequency, this variable is different for every sub-data set. Of course this procedure is done both in English and Dutch.

Figure 1a: Histogram showing the distribution of 4305 English words overs AoA.
DATASET CHARACTERISTICS

ENGLISH DATASET CHARACTERISTICS
The English base dataset contains 4305 words and their values for 11 variables. To make the content of this database more concrete, there are several examples of English words with a minimum or maximum value for every particular variable visible in Table 1. From the distribution of AoA (Figure 1a) can be concluded that children learn words in English between their second and eighteenth birthday. Furthermore the amount of words learned every year increases until the child is seven years old, where after the part of new learned words becomes smaller every year.

Figure 1b shows the distribution of words based on dominant Part of Speech (PoS). This dataset consists mainly of nouns (65.6% of all words), followed by verbs and adjectives (respectively 15.8% and 15.4%). All other words like pronouns, numerals and prepositions have the smallest portion in this dataset (3.1%).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum value</th>
<th>Examples of words with minimal values</th>
<th>Maximum value</th>
<th>Examples of words with maximal values</th>
</tr>
</thead>
<tbody>
<tr>
<td>AoA</td>
<td>2.28</td>
<td>Potty, water, wet</td>
<td>17.20</td>
<td>Binge, tenure, stipend</td>
</tr>
<tr>
<td>Number of characters</td>
<td>2</td>
<td>Do, go, ad</td>
<td>11</td>
<td>Lightweight, christening, frightening</td>
</tr>
<tr>
<td>Number of syllables</td>
<td>1</td>
<td>Act, pray, gab</td>
<td>3</td>
<td>Aspirin, separate, piano</td>
</tr>
<tr>
<td>Concreteness</td>
<td>1.12</td>
<td>Would, believe, hope</td>
<td>5.00</td>
<td>Raincoat, blender, donkey</td>
</tr>
<tr>
<td>Arousal</td>
<td>1.60</td>
<td>Grain, dull, calm</td>
<td>7.74</td>
<td>Lover, sex, gun</td>
</tr>
<tr>
<td>Valence</td>
<td>1.30</td>
<td>Rapist, torture, murder</td>
<td>8.47</td>
<td>Free, fun, happy</td>
</tr>
<tr>
<td>Dominance</td>
<td>2.14</td>
<td>Earthquake, rapist, kidnap</td>
<td>7.68</td>
<td>Smile, self, win</td>
</tr>
<tr>
<td>Lexical Decision Time (ms)</td>
<td>476.0263</td>
<td>Eye, sleep, sexy</td>
<td>917.9200</td>
<td>Muskrat, cockeyed, clothesline</td>
</tr>
<tr>
<td>CHILDES frequency</td>
<td>1</td>
<td>Unfair, gambling, depress</td>
<td>33,223</td>
<td>Do, have, one</td>
</tr>
<tr>
<td>SUBTLEX-US frequency</td>
<td>9</td>
<td>Listless, dustpan, mangle</td>
<td>314,232</td>
<td>Have, do, be</td>
</tr>
<tr>
<td>Google frequency</td>
<td>21,663</td>
<td>Curtsey, twerp, squealer</td>
<td>2,398,724,162</td>
<td>Be, have, new</td>
</tr>
<tr>
<td>Summed frequency</td>
<td>21,693</td>
<td>Curtsey, twerp, squealer</td>
<td>2,399,025,780</td>
<td>Be, have, new</td>
</tr>
</tbody>
</table>

Table 1: Minimum and maximum value of every English variable and examples of words with a minimal or maximal value.
DUTCH DATASET CHARACTERISTICS

The Dutch dataset consists of 8996 words and to make the content of this dataset more concrete the minimum and maximum values for every variable even as corresponding examples are displayed in Table 2. The distribution of these words based on AoA shows that all words are learned when children are between one and seventeen years old (Figure 2a). Furthermore the amount of words that is learned every year increases until children are ten years old, after that moment this amount is reduced every year. From the distribution of the words based on dominant Part of Speech (PoS) can be concluded that the largest part (56.2%) of all words in the database are nouns, this percentage is followed by the verbs word class and adjective word class (respectively 24.5% and 12.6%) and all other word types cover 6.7% of the complete dataset (Figure 2b).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum value</th>
<th>Words with minimum value</th>
<th>Maximum value</th>
<th>Words with maximum value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AoA</td>
<td>1.53</td>
<td>Mama, ja, papa</td>
<td>18.50</td>
<td>Tweet, Facebook, feromoon</td>
</tr>
<tr>
<td>Number of characters</td>
<td>1</td>
<td>U, ar, el</td>
<td>20</td>
<td>Communicatiemiddel, Gezichtsuitdrukking, Doorzettingsvermogen</td>
</tr>
<tr>
<td>Number of syllables</td>
<td>1</td>
<td>Pi, uk, eg</td>
<td>7</td>
<td>Aluminiumfolie, bibliothecaresse, communicatiemiddel</td>
</tr>
<tr>
<td>Concreteness</td>
<td>1</td>
<td>Waarom, minstens, de</td>
<td>5</td>
<td>e.g. Helikopter, verrekijker, Sinaasappel</td>
</tr>
<tr>
<td>Lexical Decision Time (ms)</td>
<td>333.4900</td>
<td>Tovenaar, bezoeker, rugzak</td>
<td>1453.1300</td>
<td>Marechaussee, raddraaier, sliding</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.06609195</td>
<td>Le, mi, pampa</td>
<td>1.00</td>
<td>Brugklaas, ontdekkingsreis, pinautomaat</td>
</tr>
<tr>
<td>SUBTLEX-NL frequency</td>
<td>3</td>
<td>Tafelpoot, uier, spelregel</td>
<td>1,600,888</td>
<td>Je, het, de</td>
</tr>
<tr>
<td>BasiLexYoung frequency</td>
<td>0</td>
<td>Bijvullen, levensocht, stukmaken</td>
<td>16,186</td>
<td>De, zijn, een</td>
</tr>
<tr>
<td>BasiLexAll frequency</td>
<td>8</td>
<td>Loge, enten, identiteitskaart</td>
<td>548,253</td>
<td>De, zijn, een</td>
</tr>
<tr>
<td>Summed frequency</td>
<td>13</td>
<td>Mummificeren, tegenligger, misvormen</td>
<td>1,798,195</td>
<td>Je, de, het</td>
</tr>
</tbody>
</table>

Table 1: Minimum and maximum value of every Dutch variable and examples of words with a minimal or maximal value
COMPARISON BETWEEN BOTH DATASETS

There are some interesting findings when comparing both datasets. In the first place, the Dutch dataset is twice as large as the English dataset, while the variable with the smallest amount of words in English is larger than the variable with the smallest amount of words in Dutch, respectively 13,951 values for word arousal (and validity and dominance) in English and 8,996 values for BasiLex. It seems to be that the English valence, arousal and dominance ratings contain values for many rare words. Without these variables the English dataset will be much larger. From the distribution of AoA it is visible that in Dutch words are relatively later learned than in English. In Dutch most words are learned between eight and ten years old, while in English this maximum is already when children are six years old.

Furthermore the distribution on Part of Speech is comparable for both datasets and when looking to the examples of the words, it is visible that “be” in English and its Dutch semantic equivalent “zijn” are both very frequent. Moreover it seems that the English article “the” is missing. This word is filtered out because it does not have values for arousal, valence and dominance available. Finally, the range of word frequencies are very different among the word frequency variables. The Google variable has for example its minimum and maximum between 21,663 and 2,398,724,162 while the values of the SUBTLEX-NL variable in Dutch range only from 3 till 1,600,888. This visualises also the size of the corpus where the word frequency variable is based on.

It is not possible to do a cross-linguistic comparison with these two datasets, because they do not contain exactly the same variables. But most of the variables have an equivalent in both languages, these are: Number of Characters, Number of Syllables, Concreteness, Lexical Decision Time, and the word frequency variables SUBTLEX-US and SUBTLEX-NL. Furthermore CHILDES and BasiLexYoung contain respectively spoken and written word frequencies for young children.

ANALYSES

To discover which word frequency variable best predicts AoA, two models will be fit to every dataset: Multiple Regression Analysis and Random Forests. 10-fold cross validation will be applied in every model to achieve stable outcomes and to test how reliable the results are. This means that the model is built on 90% of the data and that the remaining 10% is used to test the model. This process is repeated ten times, until every word has been one time in the test set and nine times in the training set.

MULTIPLE REGRESSION ANALYSIS

Multiple Regression Analysis (MRA) is a good analysis to show linear relationships between AoA and its predictors. This analysis generates a linear equation based on the predictor values
with which the AoA value of every new word can be predicted when the values for the predictor variables are known. For this analysis I used the software package “bootstrap” (Tibshirani & Leisch, 2017) in R (R Core Team, 2017).

RANDOM FORESTS
The Random Forests analysis consists of three steps. In the first step 500 bootstrapped regression trees are created based on random samples of the predictor variables. This collection of trees is called a forest. After that, the AoA of every single word is predicted using each tree in this forest. The last step is to combine the results across all trees to get one final predicted outcome. The software package “randomForest” (Kabacoff, 2017) will be used for this analysis in R (R Core Team, 2017). This package is based on the Random Forest algorithm from Breiman and Cutler (2015). Because this analysis is based on averaging trees, the risk of overfitting is significantly lower and there is less variance than when using a normal regression tree. Regression Trees and Random Forests are useful for untangling non-linear relationships.

RESULTS
The MRA and Random Forest analysis are performed for both the English and the Dutch dataset. In this chapter will be described what variables are significant, which relationship exists between AoA and its predictors, and to what extent the variables are good predictors of AoA using MRA or Random Forests.

ENGLISH DATASETS
MULTIPLE REGRESSION ANALYSIS
Four Multiple Regression Analyses were done to predict AoA from the variables Number of Characters, Number of Syllables, Concreteness, Arousal, Valence, Dominance, Lexical Decision Time, and Word Frequency. Each of these four analyses contains a different word frequency measure, all other variables are exactly the same. The four word frequency measures are the CHILDES frequency, the SUBTLEX-US frequency, the GOOGLE frequency and the Summed frequency. This last frequency measure is the sum of the other three frequencies. In the following will be referred to every MRA by calling the used word frequency measure.

In the CHILDES, SUBTLEX-US, Google and Summed frequency analysis was the proportion of explained variance the following: \( R^2_{\text{CHILDES}} = 0.34 \), \( R^2_{\text{SUBTLEX-US}} = 0.34 \), \( R^2_{\text{Google}} = 0.33 \), \( R^2_{\text{Summed}} = 0.33 \). The standardized regression weights of Number of characters, Number of syllables, Concreteness, Valence, Lexical decision time and Word frequency (in all four variants) are all significant, so there is an effect of these variables on AoA. The standardized regression weights

Figure 3: Graph of standardized regression coefficients with 95% confidence interval of all English variables and the CHILDES frequency.
of arousal and dominance are not significant, so there is no effect of these two variables on AoA.

The three significant variables Number of Characters, Number of Syllables and Lexical Decision Time have positive standardized regression weights. This means that an increase in the value of one of these three variables, while the values of the other variables remain unchanged, goes hand in hand with an increase in AoA. For significant variables with a negative standardized correlation coefficient this is the other way around: an increase in the value of a such a variable causes a decrease in AoA, as long as the values of all other variables in the analysis are constant. Figure 3 displays the standardized regression coefficients for the MRA with the CHILDES word frequency. These coefficients are representative for all four analyses, except for the word frequency coefficient. This coefficient was most extreme in the CHILDES and SUBTLEX-US analysis (respectively $\beta=-0.32$ and $\beta=-0.33$) and less extreme in the Google and summed frequency analyses ($\beta=-0.17$).

COMPARISON BETWEEN LEXICAL CATEGORIES
The relationships between AoA and its predictors when only looking at words of one lexical category are displayed in Figure 4. From these figures can be concluded that none of the relationships between AoA and its predictors is changed from positive to negative or the other way around when comparing the relationships based on all lexical categories with one lexical category (i.e., nouns, verbs and adjectives). This is contrary to findings in earlier research (Braginsky, Yurovsky, Marchman, & Frank, 2016; Goodman, Dale, & Li, 2008). Furthermore Lexical Decision Time has in every analysis the strongest linear relationship with AoA and the word frequency variables are in every lexical category very correlated to each other. Until an age of seven years old these word frequency variables have a negative slope (in the verbs class this slope is very steep) and after that the lines are almost horizontal. Finally, verbs and adjectives have relatively lower values than nouns and the adjectives class contains many words with a high number of syllables.
RANDOM FOREST ANALYSIS

Four Random Forest analyses were performed to predict AoA from the Number of Characters, the Number of Syllables, Concreteness, Arousal, Valence, Dominance, Lexical Decision Time and Word frequency. As before in every analysis a different word frequency measure is used (CHILDES, SUBTLEX-US, Google or Summed frequency). The proportion of explained variance was greatest in the CHILDES analysis ($R^2 = 0.55$). Followed by SUBTLEX-US ($R^2 = \ldots$)

---

*Figure 4: These graphs show the relationship between AoA and its predictors for different lexical categories in English. These lines are computed using the LOWESS smoother. This smoother is based on locally weighted polynomial regression (Cleveland, 1981).*
0.44), Google ($R^2 = 0.37$) and the summed frequency ($R^2 = 0.37$). When building the regression trees within the random forests, some variables are more important than others. In Figure 5 the importance of every variable within the analysis is visualized.

The CHILDES analysis has the highest proportion of explained variance and the CHILDES frequency is in this case also the most important predictor. In the other three analyses is Concreteness the most important variable. Furthermore Lexical Decision Time is less important is than word frequency in the CHILDES and SUBTLEX-US analysis, but more important than word frequency in the analyses using Summed and Google frequency. The

![Image of importance of variables](image_url)

**Figure 5:** This graph shows the importance of every English variable for predicting AoA using different word frequency variables in Random Forest analysis. To determine the importance of every value in every analysis, first the Mean Squared Error (MSE) is computed using the real values for every variable. Secondly, the MSE is computed again, but now using randomly permuted values for every variable. The importance of every variable is measured as the percentage of increased Mean Squared Error (%IncMSE) between these two classification errors. The greater this difference is, the more important it is to take real values and not randomly generated ones.
importance of Number of Characters, Number of Syllables, Arousal, Valence and Dominance is relatively low and comparable over all four analyses.

**DUTCH DATASETS**

**MULTIPLE REGRESSION ANALYSIS**

Four Multiple Regression Analyses were done to predict AoA from the variables Number of Characters, Number of Syllables, Concreteness, Lexical Decision Time, Prevalence and Word Frequency. Each of these four analyses contains a different word frequency measure, all other variables are exactly the same (just as in the English MRAs). The four word frequency variables are the BasiLexYoung, BasiLexAll, SUBTLEX-NL and Summed frequency. In the following will be referred to these analyses by calling the used frequency variable.

The proportion of explained variance in the BasiLexYoung, BasiLexAll, SUBTLEX-NL and Summed frequency analysis are the following:

- $R^2_{\text{BasiLexYoung}} = 0.20$,
- $R^2_{\text{BasiLexAll}} = 0.20$,
- $R^2_{\text{SUBTLEX-NL}} = 0.21$,
- $R^2_{\text{Summed}} = 0.21$.

Every variable in every analysis was significant, so there is an effect of these variables on AoA.

The magnitude of the regression coefficients from the SUBTLEX-NL analysis are visible in Figure 6. The two variables Number of Syllables and Lexical decision time have positive standardized regression weights. From this fact can be concluded that an increase in the value of one of these two variables, while all other variables remain constant, goes hand in hand with an increase in AoA. All other variables have negative standardized correlation coefficients, this means that an increase in one of these variables while the values of all other variables in the analysis remain unchanged, causes a decrease in AoA.

**COMPARISON BETWEEN LEXICAL CATEGORIES**

The standardized regression coefficients (Figure 6) show the strength of the linear relationships between AoA and its predictors for all lexical categories. In Figure 7 these relationships are visualized for different lexical categories. What stands out in the first place is that the direction of the relationship between AoA and every predictor does not reverse when selecting words of only one lexical category. This is contrary to finding in earlier research (Braginsky, Yurovsky, Marchman, & Frank, 2016; Goodman, Dale, & Li, 2008). Positive relationships stay positive and negative relationships stay negative. Furthermore the large correlation between all four frequency variables is prominent, only the last drawn red line of the summed frequency is visible and it hides the pink and purple lines of the other frequency variables. Moreover the concreteness of adjectives are relatively lower than the concreteness of other lexical categories and verbs learned after a child’s eleventh birthday have a relatively
high number of syllables. Lexical decision time has the steepest slope and is for every lexical category comparable. Prevalence has also the same kind of relationship with AoA in all lexical categories, until 10 year almost horizontal and from 10 years more negative.

**Figure 7:** These diagrams show the relationship between AoA and its predictors for different lexical categories in Dutch. These lines are computed using the LOWESS smoother. This smoother is based on locally weighted polynomial regression (Cleveland, 1981).

**RANDOM FOREST ANALYSIS**
Four Random Forest Analyses were performed to predict AoA from the Dutch predictors. Contrary to the Dutch MRAs, the proportion of explained variance did vary between the four
analyses. It was greatest for BasiLexYoung ($R^2 = 0.39$) followed by the Summed frequency ($R^2 = 0.36$) and BasiLexAll ($R^2 = 0.36$) and finally SUBTLEX-NL ($R^2 = 0.34$).

When comparing the importance of the different variables for predicting AoA, it seems that the frequency variable is most important in the analysis with the highest proportion of explained variance (BasiLexYoung), followed by Concreteness. In all other analyses this is the other way around, Concreteness is the most important variable followed by word frequency. All other variables have quite the same importance in every analysis.
CONCLUSION AND DISCUSSION

The proportion of explained variance was greater than 0.20 in all analyses, so even in Dutch as in English there exists a strong relationship between AoA and its predictors. Furthermore the regression found a number of significant predictors. From the performed MRAs can be concluded that it does not matter which word frequency variable is used to predict AoA in the MRA, because even in English as in Dutch the differences in the proportion of explained variance were very small. On the other hand, when using Random Forest analysis greater differences exist in the proportion of explained variance between the different word frequency analyses within every language. Because the proportion of explained variance is on average higher in the Random Forest analysis than in the MRA, this method is better to predict AoA in Dutch and English. A possible explanation therefore comes from the fact that MRAs are well handling linear relationships, while on the other hand Random Forests are also able to catch non-linear relationships. So the better performance of Random Forests could point to a non-linear relationship between AoA and some of its predictors.

In English the Random Forest analysis using the CHILDES word frequency and in Dutch the Random Forest Analysis using the BasiLexYoung word frequency explained the highest proportion of explained variance. This means that word frequency distributions extracted from child-directed speech or child-directed texts are the best word frequency distributions for predicting the age of acquisition, what confirms the hypothesis. In conclusion this means that the frequency distribution in which words are offered to children is a better predictor of word learning order than a word frequency distribution based on many more instances that better approaches the real distribution (i.e., Google) or a distribution based on spoken conversations between people (i.e., SUBTLEX).

A possible explanation for the fact that in the English part CHILDES frequency is a better predictor than Google frequency is that children cannot read until they are around six years old. So the only way of language input they get is from spoken language and CHILDES is based on spoken language. Furthermore Nosofsky (1987) found that paying attention to something makes people learning more effectively (Nosofsky, 1987) and children attend better to speech that is particularly directed to them and not to just some chatting at the background. So this could explain why CHILDES is a better predictor than SUBTLEX-US.

In the Dutch part it turns out that the frequency variable based on texts written for children in kindergarten and first class of elementary school (BasiLexYoung) is a better predictor than texts written for children from kindergarten to second class of high school (BasiLexAll). This is conflicting with the hypothesis. The fact that BasiLexYoung is a better predictor than SUBTLEX-NL confirms the hypothesis, an explanation therefore could be the same as in the English part, when children read or are read aloud from a book specially written for them, they attend more to the story, so learning occurs more effectively (Nosofsky, 1987).

In this study relationships between AoA and other word attributes were also considered. Firstly, a positive relationship was found between AoA and the two variables Number of Syllables and Lexical decision time. These relationships were as expected and the positive relationship between AoA and Lexical Decision Time is a confirmation of the existence of the AoA effect (Juhasz, 2005). This effect is that words acquired earlier in life are processed faster than word acquired later in life. Between AoA and Number of Characters was also a positive relationship expected and this is confirmed in the English part of this study, but a negative
effect was found in the Dutch part. The hypothesized negative relationship between AoA and Prevalence, Concreteness and Word Frequency (in all variants) was confirmed and the expected positive relationship between AoA and the three emotional variables Arousal, Valence and Dominance turned out to be false. In addition, Arousal and Dominance were not even significant. So there is no effect of these variables on AoA. Two variables had a very strong relationship with AoA, these were Concreteness and Lexical Decision Time.

FUTURE DIRECTIONS TO IMPROVE RESULTS
To improve the obtained results many actions can be taken. In the first place, quasi-objective AoA data could be used instead of subjective AoA data, because this will approach the real AoA values better. Furthermore the corpora where the used word frequency variables are relying on can be enlarged, so the distributions will better approach the real distribution. Another interesting direction is to do a comparative study between word frequency distributions based on child-directed texts (like BasLex) and word frequency distributions based on child-directed speech (like CHILDES), which one will best predict AoA? It is not possible to conclude that from this research, because the written and spoken frequency variables are not in the same language. Furthermore a cross-linguistic study could be done. This is not possible with the current datasets because the English and Dutch dataset do not contain the same variables and not all words have a semantic equivalent in both languages. Another interesting aspect is to add other predictors to discover their effect on AoA. For example semantic or orthographic relationships could be considered or a variable that shows the babiness of a word as rated by adults. Especially this last variable is interesting, because it seems that child-directed frequencies are better predictors of AoA than natural language word frequency distributions and it is possible that words with a high child-frequency are having a higher babiness value. Finally, the datasets where the AoA predictions are based on can be enlarged by obtaining variable values for new words.

REFERENCES


Brysbaert, M., Warriner, A., & Kuperman, V. (2014b). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904-911.


