CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Normative ratings for 111 Swedish nouns and corresponding picture stimuli

Tomas Lehecka

Åbo Akademi University, Tuomiokirkontori 3, 20500 Turku, Finland & Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands
Email for correspondence: tomas.lehecka@abo.fi

Abstract

Normative ratings are a means to control for the effects of confounding variables in psycholinguistic experiments. This paper introduces a new dataset of normative ratings for Swedish encompassing 111 concrete nouns and the corresponding picture stimuli in the MultiPic database (Duñabeitia et al. 2017). The norms for name agreement, category typicality, age of acquisition and subjective frequency were collected using online surveys among native speakers of the Finland-Swedish variety of Swedish. The paper discusses the inter-correlations between these variables and compares them against available ratings for other languages. In doing so, the paper argues that ratings for age of acquisition and subjective frequency collected for other languages may be applied to psycholinguistic studies on Finland-Swedish, at least with respect to concrete and highly imageable nouns. In contrast, norms for name agreement should be collected from speakers of the same language variety as represented by the subjects in the actual experiments.

1. Introduction

Psycholinguistic experiments typically employ pictorial or linguistic stimuli in order to elicit a particular type of response from subjects. The experimental tasks can either be explicitly linguistic, as in e.g. picture naming tasks, or they can be assumed to activate the language faculty more implicitly, as in e.g. categorization or similarity perception tasks often used in research on linguistic relativity effects. Producing valid analyses of psycholinguistic data presupposes that all independent and confounding variables are taken into consideration. However, due to the complex nature of language (and of communication more generally), this is far from a trivial challenge. Linguistic units, such as words or utterances, have a plethora of distinct cognitive, phonological, morphological, syntactic, semantic and pragmatic properties. To examine the effect of any one of those properties on cognition and/or linguistic processing, all the other properties need to be controlled for. Normative ratings are an invaluable tool for psycholinguistic research in that they provide a

CrossMark

means to control for properties that are not directly observable in e.g. language corpora, such as AGE OF ACQUISITION or CATEGORY TYPICALITY. In normative ratings, such properties are quantified based on the responses from a sample of subjects representing the linguistic population under investigation.

Currently, there are no large databases of normative ratings for Swedish. Ratings for certain individual lexical properties are included in some large cross-linguistic rating studies; specifically, Rofes et al. (2018) includes IMAGEABILITY ratings for 190 Swedish words, and Łuniewska et al. (2016) includes AGE OF ACQUISITION ratings for 299 Swedish words. The lack of normative ratings for Swedish prompted Blomberg & Öberg (2015) to examine whether ratings for closely related languages could be used in place of ratings collected for Swedish. Blomberg & Öberg compared the IMAGEABILITY, AGE OF ACQUISITION and FAMILIARITY ratings for 99 Swedish words (which they collected themselves) with the ratings for equivalent English words. They found that the Swedish and the English ratings were strongly correlated for IMAGEABILITY and AGE OF ACQUISITION, and moderately correlated for FAMILIARITY. Therefore, Blomberg & Öberg suggest that English ratings can be reliably transferred to Swedish for some variables, but not for all. Moreover, they note that while the correlation between the Swedish and the English ratings is strong overall, ratings for individual words may sometimes be substantially different (for variation between languages, see also Cuetos, Ellis & Alvarez 1999). As for picture stimuli, the Multilanguage Written Picture Naming Dataset (Torrance et al. 2018) includes Swedish NAME AGREEMENT norms for 260 drawings. However, these picture stimuli, taken from Rossion & Pourtois (2004), are based on the older stimuli in Snodgrass & Vanderwart (1980), which are not standardized in regard to their visual complexity nor drawing style. Therefore, modern psycholinguistic research may find it more suitable to use picture stimuli from the new MultiPic database (Duñabeitia et al. 2017) with 750 freely available and standardized drawings.

In this paper, I introduce a new dataset of normative ratings for Swedish encompassing 111 concrete nouns and the corresponding picture stimuli in the MultiPic database (Duñabeitia et al. 2017). Specifically, the dataset includes norms for NAME AGREEMENT and CATEGORY TYPICALITY regarding the picture stimuli, and norms for AGE OF ACQUISITION and SUBJECTIVE FREQUENCY regarding the corresponding words. The norms are based on responses to four separate online surveys conducted among university students who are native speakers of the Finland-Swedish variety of Swedish. By including ratings for both lexical and visual stimuli collected from the same population of subjects, the dataset is especially useful in psycholinguistic studies combining several experimental methods. Most importantly, the dataset is the first one to introduce ratings for Finland-Swedish, thus facilitating future psycholinguistic studies comparing speakers of different varieties of Swedish. Swedish is a pluricentric language with two regional standard varieties, one in Sweden (the dominant variety) and one in Finland (the non-dominant variety with about 300,000 native speakers). The main differences between the varieties concern intonation, vocabulary and idioms (Norrby et al. 2012). Furthermore, Finland-Swedish is influenced by Finnish due to a high prevalence of functional bilingualism among the Finland-Swedish population (Tandefelt 2015). This provides an interesting potential for psycholinguistic studies, as the Finnish, the Finland-Swedish, and the

bilingual populations share the same geographical and cultural space, but they speak two typologically different languages.

The aim of this paper is (i) to describe the new dataset of normative ratings for Swedish, (ii) to examine the inter-correlations between these ratings, and (iii) to examine the correlations between these ratings and those collected previously for other languages and other varieties of Swedish. The described dataset is not intended as *the* definitive source for normative ratings for all types of psycholinguistic experiments concerning (Finland-)Swedish; after all, the dataset is relatively small in terms of the number of items included in it, especially in comparison with ratings that are currently available for many of the larger (as regards the number of speakers) languages and language varieties in the world. However, it is a substantial addition to the limited array of ratings that exist for Scandinavian languages today, and it can prove useful for psycholinguistic studies that are not too broad in scope. Most importantly for the purposes of this paper, the dataset provides a solid base for a discussion on the value of having normative ratings for the specific language (variety) targeted in a psycholinguistic experiment as opposed to using ratings for typologically and genetically related languages.

2. Normative ratings in prior research

This paper presents four types of normative ratings for Swedish. Two of them concern picture stimuli in the MultiPic database (Duñabeitia et al. 2017), namely NAME AGREEMENT and CATEGORY TYPICALITY. The other two ratings concern the words that designate the objects in these pictures, namely AGE OF ACQUISITION and SUBJECTIVE FREQUENCY. The four rating types chosen for this paper constitute by no means an exhaustive set of norms for psycholinguistic studies. They do, however, represent a first step towards facilitating certain types of psycholinguistic studies with speakers of Finland-Swedish in that these rating types represent some of the most widely used ratings-based variables overall. Other such variables include e.g. IMAGEABILITY, FAMILIARITY, and IMAGE AGREEMENT. For some of those variables, existing norms from related languages can be used; e.g. IMAGEABILITY ratings have been shown to correlate strongly between Swedish and English by Blomberg & Öberg (2015) and, given that there is no reason to expect IMAGEABILITY ratings to differ among speakers of different varieties of Swedish, the current dataset does not include this variable. For other variables, collecting ratings for Swedish and its different varieties remains a task for the future. Naturally, choosing which ratings to include in analyses of experimental data depends on the specific aims and methods of each study. In what follows, I review the role that the four variables in the current dataset play in linguistic processing based on prior research.

2.1 Name agreement

NAME AGREEMENT refers to the degree of unanimity among speakers regarding what word they use as an identifier for a given object. In a typical naming task, subjects are shown pictures of objects and asked to name them. For example, when presented with a picture of a cat, most subjects respond with the word *cat* instead of other hypothetical options, such as *puss*, *kitty*, *animal*, *whiskers* etc. Pictures with a high

name agreement elicit the same word from most subjects, while pictures with a low name agreement result in varied responses. Depending on the aim of the study, one may choose to include pictures with high name agreement only (as in e.g. Gauvin et al. 2018, and Bartolozzi, Jongman & Meyer 2021) or pictures with varying levels of name agreement (as in e.g. Barry, Morrison & Ellis 1997, and Madden, Sale & Robinson 2019). When evaluating name agreement, or when using it as an independent variable in quantitative analyses, it is common to operationalize it as the H information statistic (Shannon 1948). The H index quantifies the level of name agreement as a value ≥ 0, where 0 indicates complete agreement across all participants, and increasing values reflect increasing divergence of responses. The H index is calculated using the formula $H = -\Sigma p_i \ln(p_i)$, where p_i is the proportion of participants that have given each response.

Name agreement constitutes one of the most consistently significant – and potentially confounding – variables across psycholinguistic studies involving picture stimuli. Numerous studies have shown that name agreement is highly predictive of e.g. naming latencies; pictures with high name agreement result in faster responses than pictures with low name agreement (e.g. Snodgrass & Yuditsky 1996, Barry et al. 1997, Ellis & Morrison 1998, Perret & Bonin 2018). Furthermore, name agreement needs to be considered in experiments where lemma activation is assumed to occur on a subconscious level, such as studies in linguistic relativity effects. For example, in experiments involving object categorization or visual similarity perception, it is often assumed that the same lemma is activated across all subjects in response to a given picture (e.g. Boroditsky, Schmidt & Phillips 2003, Cubelli et al. 2011). Hence, normative ratings for name agreement provide a means to control for the potential variation in subjects' (conscious or subconscious) lexical responses to picture stimuli.

From a cognitive standpoint, picture naming involves several processing stages that psycholinguistic experiments need to take into account. The first stage is OBJECT RECOGNITION, which involves the perceptual processing of the visual stimulus and the identification of its structural properties. The second stage involves the activation of the SEMANTIC REPRESENTATION of the object/concept perceived in the picture. The third stage is LEMMA SELECTION, where the activation spreads from the semantic representation of the object to the corresponding linguistic representation, i.e. a lemma. In most cases, several related lemmas are activated to varying degrees and, subsequently, the most highly activated lemma is selected for the output. Finally, the linguistic response is MOTORICALLY PLANNED to be either spoken out or written down depending on the specific task at hand. (On the cognitive processes involved, see e.g. Glaser 1992, Johnson, Paivio & Clark 1996, Levelt, Roelofs & Meyer 1999, Bonin et al. 2015).

Name agreement can vary substantially between different stimuli. Assuming that naming data is collected from a relatively homogenous group of subjects (thereby minimizing the role of confounding social variables) low name agreement is potentially caused by one of two factors. First, visual stimuli can vary as to how identifiable they are; some stimuli are more unambiguous representations of the intended object than others. Thus, discrepancies in the answers can arise at the stage of object recognition, where the structural properties of the object are encoded. Second, discrepancies can occur at the stage of lemma selection, especially if more than one

appropriate lemma is available to designate the depicted object. In such a case, several lemmas are activated to a similar degree, increasing the competition for lemma selection, which results in both longer naming latency and lower name agreement (for a more comprehensive discussion of these processes, see Vitkovitch & Tyrrell 1995, Barry et al. 1997, Cuetos et al. 1999). Naturally, in languages with considerable dialectal variation – as in Swedish (see e. g. Ivars 2015) – name agreement may be lower due to the same object having different names in different dialects. Therefore, it seems advisable that norms for name agreement be collected from the same, relatively precisely defined population as targeted in the actual experiments.

2.2 Category typicality

CATEGORY TYPICALITY (also sometimes called SEMANTIC TYPICALITY, ITEM TYPICALITY or PROTOTYPICALITY) refers to how typical or atypical an item is judged to be as a member of a given category. There is a long tradition of cognitive psychological research demonstrating that conceptual categories have graded structures, i.e. that some instances of a category are better examples (or more typical instances) of that category than others (Mervis & Rosch 1981, Medin & Smith 1984). For example, 'apple' is a very typical member of the category FRUIT, while 'pomegranate' is considerably more unusual. As a rule, typicality norms are collected from naive participants who estimate the within-category typicality of items on a seven-point scale (Battig & Montague 1969, Rosch 1975, Uyeda & Mandler 1980, Van Overschelde, Rawson & Dunlosky 2004).

Category typicality has a considerable impact on performance in many types of psycholinguistic tasks. For example, category typicality is correlated with processing performance in categorization and semantic decision tasks in both healthy (Rips, Shoben & Smith 1973, McCloskey & Glucksberg 1979, Holmes & Ellis 2006) and aphasic subjects (Kiran & Thompson 2003, Stanczak, Waters & Caplan 2006, Kiran, Ntourou & Eubank 2007). Category typicality influences also memory performance in free recall tasks in adults (Keller & Kellas 1978, Greenberg & Bjorklund 1981) and in children (Bjorklund & Thompson 1983). Moreover, subjects are likely to name typical category members prior to atypical ones when listing members of categories (Glass & Holyoak 1974, Rosch & Mervis 1975). Category typicality also influences processing speed in picture naming (Dell'acqua, Lotto & Job 2000, Holmes & Ellis 2006). Thus, typicality plays an essential part in the organization of conceptual categories. It follows, then, that for experiments involving categorization of either lexical or picture stimuli, norms for category typicality constitute a necessary tool enabling reliable analysis.

Judgments of category typicality are partially dependent on both language and culture. Interestingly, for some categories and items, typicality ratings are more correlated across languages than for others. In comparing English and Spanish speakers' typicality ratings, Schwanenflugel & Rey (1986) show that the correlations range from .94 (for BODY PARTS) to .16 (for BIRDS). Moreover, Johnson (2001) notes that typicality judgments also vary within the same linguistic community based on the respondents' level of expertise in a given subject. Thus, it seems advisable that norms for category typicality be collected from subjects representing the same population as the subjects in the given experiment.

2.3 Age of acquisition

AGE OF ACQUISITION (commonly abbreviated as AoA) represents the age at which a word is typically learned by a child. AoA norms can be derived through two distinct approaches. First, AoA norms can be obtained by examining children's production (e.g. Chalard et al. 2003, Lotto, Surian & Job 2010, Grigoriev & Oshhepkov 2013). Second, AoA norms can be obtained using estimates of adult speakers (e.g. Barca, Burani & Arduino 2002, Ferrand et al. 2008, Moors et al. 2013). While asking adults to estimate AoA is a rather indirect means to obtain AoA norms, the method has been validated by several studies comparing the two approaches. For example, Morrison, Chappell & Ellis (1997) report a correlation of .75 between subjective estimates of AoA and children's actual performance (for similar results, see also Pind et al. 2000, Schröder et al. 2012, Łuniewska et al. 2016, Brysbaert & Biemiller 2017). For this reason, subjective estimates of AoA by adults are often used as solid approximations of objective AoA in psycholinguistic studies.

AoA affects performance in a wide range of psycholinguistic tasks (for literature reviews, see e.g. Juhasz 2005, Brysbaert & Ghyselinck 2006, Johnston & Barry 2006, Łuniewska et al. 2016). For instance, words with a lower AoA are recognized faster than words with a higher AoA (Morrison, Ellis & Quinlan 1992, Holmes & Ellis 2006, Juhasz & Rayner 2006, Cortese & Khanna 2007). In particular, AoA effects are strong in lexical decision tasks and picture naming tasks, i.e. in tasks that involve the semantic system, as opposed to word naming tasks (Brysbaert, Van Wijnendaele & De Deyne 2000, Lambon Ralph & Ehsan 2006, Cortese & Khanna 2007). Crucially, Kuperman, Stadthagen-Gonzalez & Brysbaert (2012) demonstrate that AoA ratings yield a significant effect in lexical decision tasks even when controlling for other properties of the words, such as frequency, length and similarity to other words (see also Brysbaert & Cortese 2011, Ferrand et al. 2011).

Several explanations have been proposed for the effects of AoA on linguistic processing. Some scholars have argued that words' AoA ratings reflect their cumulative frequency, i.e. the frequency with which subjects have been exposed to these words over their entire lifetime (Lewis 1999, Lewis, Gerhand & Ellis 2001). According to this view, AoA ratings are a valuable complement to so-called 'objective' corpus frequencies as predictors in psycholinguistic experiments because most available corpora consist of texts written for adult readers (meaning that the language we hear during the first 10 years of our lives is underrepresented in such corpora; see also Zevin & Seidenberg 2002). However, it is also possible that the AoA effect is, at least partly, independent of word frequency, and that it reflects an advantage in the accessibility of earlier learned words in the semantic networks (Brysbaert, Van Vijnendaele & De Deyne 2000, Sailor, Zimmerman & Sanders 2011). According to Ellis & Lambon Ralph (2000), this explanation is supported by the network plasticity hypothesis, postulating that plasticity diminishes with age, giving a processing advantage to earlier introduced items (see also Monaghan & Ellis 2010).

AoA norms have been shown to be relatively highly correlated across languages. In a large study encompassing 25 languages (including Swedish), Łuniewska et al. (2016) demonstrate significant correlations across all language pairs. For example, the Spearman's rank correlation coefficients for those language pairs that include

Swedish range from .71 to .91. Similarly, Blomberg & Öberg (2015) report a strong (.82) correlation between Swedish and English AoA ratings. However, Łuniewska et al. (2016) also show that estimates of AoA vary within language groups based on whether the raters have children or whether they are bilingual. In contrast, they show no effect of age, gender or education on AoA estimates (see Birchenough, Davies & Connelly 2017). One of the aims of the current paper is to compare the strength of the correlation between AoA estimates for two varieties of the same language (Finland-Swedish and Sweden-Swedish) against those between different languages.

One potential source of variation when comparing different AoA norms may be that different databases have been collected using different response scales, ranging from five-point Likert-type responses (e.g. Alario & Ferrand 1999, Tsaparina, Bonin & Méot 2011) to asking subjects to write down the actual age in years (e.g. Kuperman et al. 2012, Łuniewska et al. 2016). In addition, there may be some variation between datasets regarding the wording of the question used to elicit the ratings. Among those studies that report the exact form of the question, the most frequent wording concerns the subjects' own experience ('When do you think you learned this word?'). However, based on a one-language control study of question wording effects, Łuniewska et al. (2016) propose that a more suitable question concerns children in general ('When do children learn the word . . . ?'). The latter wording was used for collecting the AoA ratings in the current dataset.

2.4 Subjective frequency

SUBJECTIVE FREQUENCY refers to the subjective estimates by native speakers regarding the frequency with which a word is used in daily communication. Subjective frequency ratings are typically acquired by asking speakers to estimate word frequencies on a Likert scale. Thus, unlike frequency counts obtained from corpora, subjective frequency estimates are based solely on speakers' intuitions.

Word frequency is the most studied independent variable in psycholinguistics. In fact, frequency effects permeate a wide range (if not all) linguistic processes (for literature reviews, see e.g. Ellis 2002, Ambridge et al. 2015). With respect to lexical processing in adults, frequent words are comprehended and named faster than infrequent words (e.g. Postman & Conger 1954, Jescheniak & Levelt 1994). Most probably, these effects arise at the level of phonological encoding, where words have different activation thresholds depending on their frequency; hence, the time taken for activation and selection of a target word in the mental lexicon is shorter for high-frequency words than for low-frequency words (Jescheniak & Levelt 1994, Barry et al. 1997, Levelt et al. 1999).

Both subjective and objective frequency measures have been used as predictors in analyses of experimental psycholinguistic data. Interestingly, some older studies found that subjective frequency is a more significant predictor of visual and auditory word processing than corpus-based frequency counts (Connine et al. 1990, Balota, Pilotti & Cortese 2001). A potential explanation for this is that many of the most widely used large corpora have been dominated by relatively formal written language (e.g. newspaper text), which does not represent the type of language that people most commonly encounter. Hence, some scholars have argued that so-called

objective frequency counts suffer from a sampling bias (e.g. Gilhooly & Logie 1980, Gernsbacher 1984; recall discussion on estimates of AoA above). In order to circumvent this problem, newer psycholinguistic studies have been using subtitle corpora as sources for word frequency measures. Indeed, subtitle-based word frequency measures have been shown to outperform traditional written language corpora when predicting performance in many psycholinguistic tasks and across a wide spectrum of languages (e.g. Brysbaert & New 2009, Cai & Brysbaert 2010, Dimitropoulou et al. 2010, Boada et al. 2020). Furthermore, frequency measures based on Facebook and Twitter corpora have been shown to have equally good predictive power as subtitle-based measures, at least for some types of tasks (Herdağdelen & Marelli 2017). While subtitle (and social media) corpora provide a reliable measure of word frequency in daily communication, i.e. of the so-called OBJECTIVE FREQUENCY, subjective frequency estimates do still constitute a relevant variable for some types of studies. According to Brysbaert & Cortese (2011), subjective frequency estimates may contribute significantly to analyses of psycholinguistic data in instances where the ratings and the experimental data have been collected from the same population, and where this population is underrepresented in the available language corpora (e.g. Finland-Swedish university students). Furthermore, subjective frequency estimates are useful for investigating linguistic intuitions.

Subjective frequency estimates have been shown to be highly correlated with objective frequency counts from corpora (Balota et al. 2001). The exact degree of this correlation depends partly on what corpora are used to assess the 'objective' frequency (see above) and partly on how the subjective frequency estimates are collected. Balota et al. (2001) show that subjective frequency estimates collected online are consistently higher than estimates collected with printed surveys. Nonetheless, the estimates are highly correlated across the different collection methods (> .9). Contrary to what might be expected, Balota et al. (2001) do not find any effect of age on the subjective frequency estimates, but they do show that estimates vary slightly when subjects are asked to estimate word frequencies in different domains (e.g. 'written' vs. 'heard' language). Typically, subjective frequency estimates have been obtained by asking how often participants, on average, 'encounter' a given word (i.e. encompassing all language domains). Responses are normally collected on a seven-point scale ranging from 'several times a day' to 'almost never' (e.g. Balota et al. 2001, Boukadi, Zouaidi & Wilson 2016, Soares et al. 2017).

3. Method

The picture stimuli used in this study constitute a subset of the MultiPic dataset (Duñabeitia et al. 2017). The MultiPic dataset is the result of a collaborative European project intended to provide cognitive scientists with a publicly available and standardized set of stimuli for linguistic experiments. The dataset consists of 750 relatively simple drawings of concrete concepts. All drawings are made in the same graphic style (see examples in Figure 1). The entire dataset has been normed with respect to name agreement and visual complexity for six languages (so far): British English, Dutch (separately for Belgium and the Netherlands),
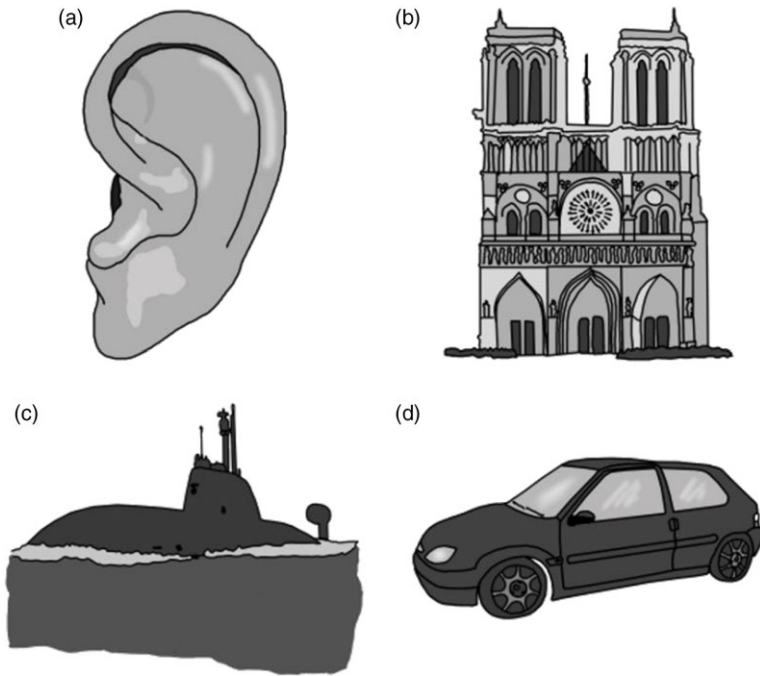
Figure 1. Examples of picture stimuli from the MultiPic database.

French, German, Italian, and Spanish. Grayscale and colored versions of the picture stimuli, as well as the norms for each language, are available at www.bcbl.eu/databases/multipic/.

I selected a subset of 111 picture stimuli from the MultiPic dataset based on the following criteria: (i) the stimuli should represent a variety of semantic categories, (ii) each semantic category in the subset should include at least nine members (i.e. nine different picture stimuli), (iii) the members of each semantic category should intuitively represent varying levels of category typicality, (iv) the expected Swedish names for the picture stimuli within each semantic category should be as equally distributed between the two grammatical genders as possible. Hence, the selected subset includes a relatively varied group of stimuli, while controlling for certain semantic (category membership) and grammatical (gender) variables. This, in turn, enables examining the effects of those variables in psycholinguistic experiments using this relatively limited number of stimuli. For example, the stimuli may be used to investigate the effects of grammatical gender on object categorization and similarity perception, which have been studied extensively in recent years in many languages (for a literature review of this line of research, see Samuel, Cole & Eacott 2019).

The selected stimuli represent the following semantic categories: ANIMAL (14 pictures), BODY PART (16), BUILDING (10), CLOTHING (15), FRUIT (10), HOME (13), NATURE (13), MUSICAL INSTRUMENT (9), and VEHICLE (11). I chose to use the grayscale version of the stimuli for two main reasons. First, colors add an additional variable to visual processing, where some colors draw more attention than others

(e.g. Elliot & Maier 2014, Kuhbandner et al. 2015). For some types of cognitive tasks, especially similarity perception or similarity judgment tasks, the use of color may therefore affect the responses. Moreover, given that many (if not most) objects do not have a 'natural' color (e.g. it is equally 'natural' for a car to be red as it is for it to be blue), color can in some cases constitute an unnecessary confounding variable – and one that is difficult to operationalize quantitatively in e.g. regression analyses. Using grayscale pictures eliminates this source of variability between the stimuli. Second, given that the data collection was carried out online and that colors can look quite different on different electronic devices, using grayscale drawings reduces the effect of display variation.

I collected the data through four separate online surveys (one for each rating type). All respondents were students at the Åbo Akademi University in Turku, Finland. Most of the students at the university speak Swedish as their first language. The majority of the students come from the coastal regions of Finland, with the three largest regions Uusimaa (Swedish: Nyland), Southwest Finland (Egentliga Finland) and Ostrobothnia (Österbotten) each accounting for roughly 30 % of the students (based on the survey data). I recruited the subjects by sending out survey invitation emails to student mailing lists. Depending on the rating type, I stressed different inclusion criteria in the invitations (see below). As participation was anonymous, it cannot be ruled out that some subjects participated in multiple surveys. However, given that the surveys were conducted at separate times, and that the surveys consisted of different tasks, it is unlikely that possible multiple participation could affect the responses. Participation in the surveys was voluntary, and the subjects did not receive any compensation for their participation.

The inclusion of subjects in the finalized datasets for each rating type was based on the subjects' responses to questions regarding their background, and on metrics pertaining to the dispersion and the speed of their responses. Specifically, participants were excluded if their ratings (for category typicality, AoA, or subjective frequency) had a standard deviation of less than 1. Such lack of dispersion in the responses by a participant, i.e. giving almost every item an identical rating, suggested that they had not been attentive to the task at hand. In addition, participants were excluded if they had performed the rating tasks faster than what was deemed realistic for completing them in an attentive manner. Based on a pilot run of the surveys with five participants who were asked to complete the survey as fast as possible while still being attentive to the task at hand, a minimum time limit of 40 seconds per survey page was established (meaning, in effect, spending less than 1.8 seconds per item). None of the five participants in the pilot had completed any survey page in less than 49 seconds. Participants who spent less than 40 seconds per survey page were therefore excluded from the analyses.

Information about the subjects' linguistic background was collected by means of three questions: (i) 'What is your first language? You may indicate two languages if you consider yourself a balanced bilingual.', (ii) 'What is your proficiency in Swedish?', and (iii) 'What languages do you have at least a working proficiency in?'. In this paper, I use the term MONOLINGUAL of subjects who indicate that Swedish is their ONLY first language and who consider to have native proficiency in Swedish. In contrast, I use the term BILINGUAL of subjects who indicate that both Swedish and Finnish are their first languages. Thus, I use the terms solely to refer to

the number of L1s a subject has. In reality, virtually all subjects are proficient in English and Finnish, and many of them speak even more languages. As with most volunteer-based studies, the gender distribution is heavily skewed towards female subjects (Smith 2008). While this can be seen as somewhat unfortunate, it is also likely that any future psycholinguistic studies of Finland-Swedish volunteers will face a similar gender distribution. I built the surveys and collected the data using the SoSciSurvey online platform (Leiner 2019). First, I collected the data for the ratings regarding the picture stimuli (name agreement and category typicality), and afterwards for ratings regarding the corresponding words (age of acquisition and subjective frequency). The full set of normative ratings is available in online Supplementary materials to this paper.

3.1 Survey 1: Name agreement

Forty-eight native speakers of Swedish participated in the survey (16 of whom were Swedish–Finnish bilingual; 39 female, seven male, two other; median age group = 20–22 years). The subjects saw the complete set of 111 pictures one at a time on their computer/tablet screen. They were asked to write down the word (and only one word) which they intuitively think is the best representation of the depicted object. The pictures were presented in a randomized order. All pictures had the dimensions 130 pixels × 130 pixels. The task was divided into five sets of 22/23 pictures and the subjects were encouraged to take short breaks between each set. There was no time limit to the task. On average the subjects completed the task in 12 minutes. For each picture item, the final norms include the MODAL response (i.e. the word given by the majority of the subjects), the number and the proportion of modal responses, the number of distinct responses (i.e. answers given by at least one subject), the number of idiosyncratic responses (i.e. answers given by no more than one subject), the number of invalid responses, and the H index for name agreement. If $H = 0$, then all subjects have given the same response to a given stimulus. Consequently, the value of H increases as a function of divergence among the responses.

3.2 Survey 2: Category typicality

The procedure of displaying the picture stimuli in Survey 2 was identical to that in Survey 1. Each subject saw the complete set of 111 drawings. However, instead of naming the object in the drawing, the subjects in Survey 2 were asked to rate the prototypicality of the object as a member of a given semantic category. For example, the subjects saw a drawing of an apple, with the following text directly beneath the drawing: 'How prototypical is the depicted object as a member of the category FRUIT?' (Swedish: 'Hur prototypisk medlem av kategorin FRUKT är objektet på bilden?'). The subjects gave their responses on a seven-point scale directly under the question. The scale was labeled with numbers from 1 to 7. The extreme values were also labeled lexically with 1 = 'Very atypical' and 7 = 'Very prototypical', respectively.

Fifty-five subjects took part in the survey. Seven subjects were excluded from the analysis because they indicated that they were not native speakers of Swedish. Another seven subjects were excluded due to their ratings showing extremely low levels of dispersion across items (SD < 1), which suggests that they were not

attentive to the task at hand. Of the 41 subjects included in the analysis, 34 were female, six male and one other (mean age = 20.2 years). Ten of the subjects were Swedish–Finnish bilinguals. On average, the subjects completed the task in nine minutes. The final category typicality ratings include for each drawing the mean rating from all 41 subjects together with the standard deviation of the ratings.

3.3 Survey 3: Age of acquisition

The stimuli for the Surveys 3 and 4 consisted of the modal responses to the picture naming task in Survey 1. The subjects were presented with the complete set of 111 words divided into five pages with 22/23 words each. The words were presented in random order. The participants were asked to estimate the AoA for each word. Following the question wording proposed by Łuniewska et al. 2016, the responses were elicited using the question: 'At what age do children speaking Swedish as L1 typically learn the word …?'. In accordance with Stadthagen-Gonzalez & Davis (2006), and Kuperman et al. (2012), I specified AoA as 'the age at which children understand the meaning of the word' beneath the question. Following Gilhooly & Logie (1980) and Soares et al. (2017), the subjects responded on a seven-point scale labeled as follows: 1 = '0–2 years', 2 = '3–4 years', 3 = '5–6 years', 4 = '7–8 years', 5 = '9–10 years', 6 = '11–12 years', and 7 = '13 or older'.

Only monolingual native Swedish speakers were invited to participate in the survey. Fifty-two subjects participated in the survey and two of them were excluded from the analysis because they indicated that they were not monolingual native speakers of Swedish. In addition, six subjects were excluded from the analysis due to having completed the task quicker than was deemed realistic based on test runs of the survey (less than 40 seconds per page). Hence, the final rating data comprises the responses from 44 subjects. On average, the subjects completed the task in 10 minutes. The final dataset includes the mean AoA estimate for each word on a scale from 1 to 7 and the standard deviation of the estimates.

3.4 Survey 4: Subjective frequency

In Survey 4, subjects were asked to estimate the frequency of the 111 words representing the modal responses to the picture naming task in Survey 1. Similarly to Survey 3, the words in Survey 4 were presented in random order and divided into five sets with 22/23 words each. The subjects were asked to estimate how often they think that they 'hear/say/read/write the following words'. For each word, the subjects were asked to pick the most appropriate alternative from a seven-point scale labeled as follows (following Balota et al. 2001): 1 = 'Almost never', 2 = 'Once a year', 3 = 'Once a month', 4 = 'Once a week', 5 = 'Three or four times a week', 6 = 'Once a day', and 7 = 'Several times a day'.

Fifty-one subjects took part in the survey. Only monolingual native Swedish speakers were invited to participate. Four subjects were excluded from the data because they indicated that they are not monolingual Swedish speakers. Additional 4 subjects were excluded on the grounds of having completed the task quicker than was deemed realistic (< 40 s per page). Thus, the final ratings comprise the responses from 43 subjects. On average, the subjects completed the task in 14

minutes. The final dataset includes the mean subjective frequency estimate for each word on a scale from 1 to 7 and the standard deviation of the estimates.

4. Results

4.1 Summary statistics

Table 1 presents the summary statistics for all four variables investigated in the surveys: name agreement, category typicality, age of acquisition and subjective frequency. The table includes two measures for name agreement, namely, the percentage of subjects who gave the modal response, and the *H* index. The two measures are included here because the percentage constitutes a more easily interpretable number, while the *H* index is a more specific measure for the divergence of the responses. A reliability analysis of the three ratings-based variables (category typicality, AoA and subjective frequency) indicates high inter-rater correlations, which strongly supports the reliability of the data: the intra-class coefficient (ICC) for category typicality (using a two-way random effect model based on the absolute agreement of multiple raters): M = .98, with 95% confidence interval (CI): .96–.98; ICC for AoA: M = .97, CI: .97–.98; ICC for subjective frequency: M = .98, CI: .98–.99.

Table 1 demonstrates that across all 111 picture stimuli, the modal responses to the picture naming task constitute 83 percent of the total number of responses. However, for most stimuli, the level of name agreement is higher, as demonstrated by the median value of the percentage of modal responses (.92). For 12 of the 111 stimuli, the subjects were completely unanimous in their response. For example, the picture in Figure 1a elicited the word *öra* 'ear', from all subjects. In contrast, the picture in Figure 1b elicited many different responses, resulting in the lowest level

Table 1. Summary statistics for ratings for name agreement, category typicality, age of acquisition and subjective frequency.

	Name agreement		Typicality	AoA	Subj. freq.
	% Mod. response	*H*-index			
Mean	0.83	0.57	4.68	2.60	3.93
SD	0.19	0.47	1.22	0.83	1.23
Median	0.92	0.47	4.78	2.48	3.74
Min	0.31	0.00	2.02	1.36	1.79
Max	1.00	1.81	6.90	5.27	6.40
Q1	0.73	0.19	3.67	2.05	2.93
Q3	0.98	0.90	5.56	3.15	5.01
IQR	0.25	0.71	1.89	1.10	2.08
Skew	0.33	1.48	0.70	1.55	1.56

H index: increasing values indicate decreasing name agreement; SD = standard deviation; Min = minimum value across all items; Max = maximum value across all items; Q1 = 25th percentile; Q3 = 75th percentile; IQR = inter-quartile range; Skew = (Q3 – Median)/(Median – Q1) (value larger than 1 means the distribution is positively skewed).

of name agreement among all the stimuli (31% modal answers, modal response: *katedral* 'cathedral', $H = 1.81$).

As for the other three variables, Table 1 shows that the subjects used the entire seven-point scale in their responses regarding category typicality and subjective frequency. The lowest mean rating for category typicality (2.02) concerns Figure 1c and the highest (6.92) Figure 1d (both represent the semantic category VEHICLE). The lowest mean rating for subjective frequency concerns the word *tamburin* 'tambourine' (1.79) and the highest the word *säng* 'bed' (6.40). Furthermore, the means of the category typicality and subjective frequency ratings lie approximately around the midpoint of the seven-point scale (4.68 and 3.93, respectively). In contrast, most of the ratings for AoA lie toward the lower end of the scale, i.e. most of the words are judged to be acquired relatively early (mean rating 2.60). Moreover, both the standard deviation (0.83) and the inter-quartile range (1.10) of the ratings are substantially smaller for AoA than for either category typicality or subjective frequency. This suggests that the words in the current dataset demonstrate relatively little variation in terms of their estimated AoA. The lowest mean rating for AoA concerns the word *hand* 'hand' (1.36) and the highest the word *katedral* 'cathedral' (5.27).

4.2 Correlations

In this section, I examine the relationships between different ratings. First, I discuss the inter-correlations between the four rating types collected for the current dataset. Thereafter, I compare the current ratings for name agreement, AoA and subjective frequency against ratings acquired in prior studies. There are no existing ratings for category typicality regarding the MultiPic stimuli, which is why no comparisons can be made concerning this variable.

4.2.1 Inter-correlations

Table 2 shows the Spearman's rank correlation coefficients between the H index, and the ratings for category typicality, AoA and subjective frequency in the current dataset. The table demonstrates that there is a weak negative correlation between the H index and category typicality ($-.22$), indicating that higher typicality ratings generally coincide with somewhat higher name agreement (i.e. lower H index). In other

Table 2. Inter-correlations (Spearman's rank correlation) between the four rating values (H index, mean Category typicality, mean AoA and mean Subjective frequency).

	H index	Typicality	AoA	Subj. frequency
H index	1			
Categ. typicality	−.22*	1		
AoA	.30**	−.45***	1	
Subj. Frequency	−.21*	.49***	−.66***	1

* $p < .05$; ** $p < .01$; *** $p < .001$

words, pictures of objects which are judged to be prototypical members of a semantic category elicit relatively uniform responses in the naming task, while atypical objects elicit somewhat more varying responses from the subjects. There is also a weak negative correlation between the H index and subjective frequency ($-.21$), meaning that the modal responses to picture stimuli with high name agreement ratings (i.e. low H index) represent words which are estimated to be relatively frequent in Swedish overall. On the other hand, there is a weak positive correlation (.30) between the H index and AoA, suggesting that picture stimuli with high name agreement ratings (i.e. low H index) tend to elicit words that are estimated to be learned early in childhood (i.e. low AoA). There is also a moderate negative correlation between the category typicality of the objects in the picture stimuli and the AoA of the corresponding words ($-.45$). In other words, objects which are perceived as typical members of a semantic category (i.e. high category typicality) elicit words which are judged to be learned relatively early in childhood (i.e. low AoA). Finally, Table 2 shows that the two highest correlation coefficients concern the link between subjective frequency and category typicality, and subjective frequency and AoA, respectively. Subjective frequency is positively correlated with category typicality (.49), meaning that pictures of objects which are perceived as typical representatives of a semantic category elicit words which are estimated to have a high frequency in Swedish. In contrast, subjective frequency is strongly negatively correlated with AoA ($-.66$), suggesting that, as a rule, words which are estimated to have a high frequency in language use are also estimated to be learned relatively early in childhood.

Below, I compare the ratings for name agreement, AoA and subjective frequency in the current dataset against corresponding ratings acquired in prior research for Swedish and for other languages. Additionally, I compare the ratings for subjective frequency against objective corpus-based measures of lexical frequency in Swedish. Unfortunately, there are no existing ratings for category typicality concerning the stimuli used in the current dataset (i.e. the MultiPic database), making comparisons for this variable impossible.

4.2.2 Name agreement

Table 3 shows the Spearman's rank correlation coefficients between the name agreement ratings (H index) collected for the current dataset and ratings acquired in identical naming tasks for British English, Dutch (Belgium), Dutch (Netherlands), French, German, Italian, and Spanish, respectively, in Duñabeitia et al. (2017). All the ratings have been collected using picture stimuli in the MultiPic database. I calculated the correlations for the 111 picture stimuli included in the current dataset. The correlations between name agreement ratings are significant for all language pairs. The correlations for the language pairs that include Swedish range from .35 to .53 (i.e. weak to moderate), while the correlations for the other language pairs range from .38 to .71. Thus, the correlations between the Swedish data and the data in Duñabeitia et al. (2017) are in the same range as those found within the MultiPic database, albeit in the lower end of that spectrum. I discuss the implications of these findings in the Discussion section of the paper.

Table 3. Spearman's rank correlation coefficients regarding name agreement ratings (*H* index) in the current dataset (Swedish) and ratings for the languages in the MultiPic database (Duñabeitia et al. 2017).

	Swedish	British English	Dutch (BE)	Dutch (NL)	French	German	Italian	Spanish
Swedish	1							
British English	.42***	1						
Dutch (BE)	.53***	.47***	1					
Dutch (NL)	.51***	.55***	.71***	1				
French	.50***	.51***	.58***	.65***	1			
German	.48***	.56***	.55***	.64***	.54***	1		
Italian	.35***	.38***	.44***	.41***	.53***	.54***	1	
Spanish	.37***	.54***	.51***	.54***	.56***	.46***	.58***	1

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 4. Correlations (Spearman's rank correlation) between AoA ratings in the current dataset (Lehecka (Swedish)) and ratings in four other databases: Brysbaert & Biemiller (2017; English), Kuperman, Stadthagen-Gonzalez & Brysbaert (2012; English), Lind et al. (2015; Norwegian), and Łuniewska et al. (2016; Swedish).

	Lehecka (Swedish)	B&B (English)	Kuperman et al. (English)	Lind et al. (Norwegian)	Łuniewska et al. (Swedish)
Lehecka (Swedish)	1				
B & B (English)	.79***	1			
Kuperman et al. (English)	.77***	.89***	1		
Lind et al. (Norwegian)	.85***	.68***	.65***	1	
Łuniewska et al. (Swedish)	.79***	.70***	.60***	.71***	1

* $p < .05$; ** $p < .01$; *** $p < .001$

4.2.3 Age of acquisition

Table 4 demonstrates the correlations between the AoA ratings collected for Swedish in the current dataset (Lehecka (Swedish)) and AoA ratings reported in four prior studies. Brysbaert & Biemiller (2017) is currently the largest dataset of AoA ratings for any language; it contains 44,000 English words. Moreover, it is the only dataset in Table 4 that contains test-based AoA ratings (i.e. the ratings are based on word recognition tests administered to children of different ages). The other datasets in the table contain AoA ratings based on estimates made by adult speakers. I used the modal responses in the British English dataset of the MultiPic database to identify corresponding English words to the 111 Swedish

nouns in my dataset. I then searched for these English words in the datasets of Brysbaert & Biemiller (2017) and Kuperman et al. (2012). The former includes ratings for 108 and the latter for 106 of the 111 words. The correlation between the subjective AoA ratings in my dataset and both the objective and subjective AoA ratings for English are strong (.79 and .77, respectively). The third dataset used for comparison, Lind et al. (2015), contains AoA ratings for 1,650 words in Norwegian. It is the largest database of normative ratings in any North Germanic language. It includes 78 of the 111 words in my dataset. There is a strong correlation (.85) between the AoA ratings in Lind et al. (2015) and the current dataset. Finally, Łuniewska et al. (2016) contains AoA ratings for 299 words in 25 languages, including Swedish (collected in Sweden). Their dataset includes 49 of the 111 words in my dataset. The ratings in the two datasets of Swedish words are strongly correlated (.79). Interestingly, however, Table 4 shows that there is an equally strong correlation between the AoA ratings collected from two different populations of Swedish speakers as there is between ratings collected from speakers of different Germanic languages. I discuss the implications of this finding in the Discussion section of the paper.

4.2.4 Subjective frequency

There are relatively few large datasets of subjective frequency ratings available for any language. Balota et al. (2001) contains subjective frequency estimates for nearly 3,000 English monosyllabic words, Soares et al. (2017) for 3,800 Portuguese words, and Desrochers & Thompson (2009) and Ferrand et al. (2008) for 3,600 and 1,500 French words, respectively. Table 5 illustrates the correlations (Spearman's rank correlation coefficients) between the subjective frequency ratings in the current study and the ratings for the corresponding words in the abovementioned English, Portuguese and French databases. I used the English translations in Soares et al. (2017) and the modal responses in the British English and the French datasets in the MultiPic database (Duñabeitia et al. 2017) to determine these correspondences. Balota et al. (2001) contains ratings for 56, Desrochers & Thompson (2009) for 75, Ferrand et al. (2008) for 39, and Soares et al. (2017) for 86 of the 111 items in the current dataset. Table 5 demonstrates that the correlations between the subjective frequency estimates are high for all language pairs, ranging from .71 to .87.

It is also informative to compare the subjective frequency estimates in the current dataset against actual corpus-based frequencies. Table 6 demonstrates the correlations (Spearman's rank correlation coefficients) between the subjective frequency estimates and lemma frequencies in ten separate Swedish corpora in the Språkbanken database (Borin, Forsberg & Roxendal 2012) as well as the frequencies extracted from the OpenSubtitles corpus (specifically, the subcorpus of Swedish subtitles for English movies and TV series; Lison & Tiedemann 2016). The corpora in the table represent different genres and regional varieties. The table shows that the strength of the correlation varies considerably depending on which corpus the subjective frequency estimates are compared to. The subjective estimates are most strongly correlated with lemma frequencies in blog text corpora (Bloggtexter 2006–2013, Bloggmix 2011). They are also strongly correlated with frequencies

Table 5. Correlations (Spearman's rank correlation) between subjective frequency ratings in the current dataset (Lehecka (Swedish)) and ratings in four other databases: Balota, Pilotti & Cortese (2001; English), Desrochers & Thompson (2009; French), Ferrand et al. (2008; French), and Soares et al. (2017; Portuguese).

	Lehecka (Swedish)	Balota et al. (English)	D&T (French)	Ferrand et al. (French)	Soares et al. (Portuguese)
Lehecka (Swedish)	1				
Balota et al. (English)	.78***	1			
D&T (French)	.78***	.71***	1		
Ferrand et al. (French)	.71***	.87***	.83***	1	
Soares et al. (Portuguese)	.85***	.84***	.86***	.85***	1

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 6. Spearman's rank correlation coefficients between subjective frequency estimates in the current dataset and lemma frequencies in 10 Swedish corpora in Språkbanken as well as the frequencies in the OpenSubtitles corpus (Lison & Tiedemann 2016).

Corpus	r_s
Bloggtexter 2006–2013 (Finland-Swedish)	.78***
Bloggmix 2011	.77***
Barnlitteratur 1988–2013 (Finland-Swedish)	.66***
Skönlitteratur 2000–2013 (Finland-Swedish)	.58***
Twitter 2017	.54***
LäSBarT: Lättläst svenska och barnbokstext	.49***
Ungdomslitteratur 1992–2011 (Finland-Swedish)	.49***
PAROLE	.48***
OpenSubtitles_Swedish	.46***
Hufvudstadsbladet 2014 (Finland-Swedish)	.34***
Åbo Underrättelser 2013 (Finland-Swedish)	.28**

* $p < .05$; ** $p < .01$; *** $p < .001$

in Finland-Swedish children's books (Barnlitteratur 1988–2013) and Finland-Swedish fiction (Skönlitteratur 2000–2013). In contrast, they are only weakly correlated with lemma frequencies in Finland-Swedish newspaper corpora (Åbo Underrättelser 2013, Hufvudstadsbladet 2014). Thus, the subjects' intuitions regarding the frequencies of concrete nouns in Swedish seem to correspond well to actual language use in blogs – perhaps the most informal of the genres represented in Table 5. Whether this genre actually represents the most frequent type of language that the subjects encounter, or whether the correspondence is due to some other

factor, lies beyond the scope of this paper. Interestingly, the subjective frequency estimates correlate less strongly with the frequencies in the OpenSubtitles corpus (.46). This is noteworthy given that word frequencies in subtitle corpora have been shown to constitute an accurate predictor of performance in many psycholinguistic tasks (see Section 2.4). Furthermore, the subjective frequency estimates correlate equally weakly with the frequencies in the PAROLE corpus, which was originally designed as a balanced corpus representing the use of contemporary language (for a general overview of the PAROLE project, see e.g. Marinelli et al. 2003). I discuss the implications of these findings in the Discussion.

5. Discussion

The dataset introduced in this paper constitutes a substantial addition to the currently available normative ratings for Swedish. In particular, the dataset is the first one to provide ratings for the Finland-Swedish variety of Swedish. The dataset includes norms for name agreement and category typicality for 111 picture stimuli from the MultiPic database (Duñabeitia et al. 2017), and norms for age of acquisition and subjective frequency for 111 Swedish nouns that correspond to the picture stimuli. Given that all four rating types are collected from the same population of subjects (Swedish speaking students at the Åbo Akademi University in Finland), the combination of ratings is especially useful in studies employing multivariate analyses. While the set of 111 stimuli is far from large, the stimuli do represent a variety of semantic categories as well as different grammatical genders. Hence, the ratings may be used for studies involving e.g. categorization tasks, similarity perception tasks or lexical decision tasks.

Examining inter-correlations between the four rating types reveals that there is only a relatively weak correlation between name agreement (H index) and the estimates for category typicality, AoA, and subjective frequency, respectively. This suggests that name agreement, as a psycholinguistic variable, is relatively independent from the other variables in the dataset. In contrast, the inter-correlations between category typicality, AoA and subjective frequency are moderate to strong. The strongest correlation is found between AoA and subjective frequency ratings, where words that are thought to be acquired at a younger age are consistently estimated to be more frequent in language overall. This brings into question whether AoA estimates are, indeed, based on intuitions regarding word acquisition or whether subjects use word frequency as a heuristic for such estimates (and vice versa) as some scholars have suggested (Lété & Bonin 2013, Baayen, Milin & Ramscar 2016). Investigating this question lies beyond the scope of this paper. Nevertheless, the strong correlation (>.7) between subjective AoA estimates in my dataset and BOTH the subjective AND objective AoA ratings for English (Kuperman et al. 2012, Brysbaert & Biemiller 2017) suggests that the subjective AoA estimates can be considered reasonably accurate.

The current paper shows that normative ratings collected for other languages may be applied to Finland-Swedish in regard to some rating types, but not all. In accordance with Blomberg & Öberg (2015) and Łuniewska et al. (2016), the paper demonstrates that AoA ratings for Swedish are highly correlated with those for

other languages. Thus, the results support the notion that AoA norms from closely related languages (e.g. Kuperman et al. 2012, Lind et al. 2015, Brysbaert & Biemiller 2017) lend themselves to be adopted in psycholinguistic studies on Swedish. Naturally, cross-linguistic use of AoA norms should only be applied to unambiguous words that are easily translatable. Importantly, the results in the current study demonstrate that the correlation between AoA ratings collected from two groups of Swedish speakers (the current dataset and Łuniewska et al. 2016) is not any higher than the correlation between Swedish and English or Swedish and Norwegian AoA ratings. Hence, the amount of variation in AoA estimates is equally large across different speaker groups of the same language as it is across speakers of closely related languages. In slightly hyperbolic terms, if the AoA ratings are not collected from the same specific population as the experimental data is (e.g. Finland-Swedish university students), then one can just as well use ratings from a different language.

Like AoA ratings, also subjective frequency estimates seem at first sight a good candidate for cross-linguistic use. The results in the current study demonstrate a strong correlation between the subjective frequency ratings for the 111 Swedish nouns and ratings previously collected for English, French and Portuguese. However, one should keep in mind that the words in the current dataset only contain concrete and highly imageable nouns. It is possible that subjective frequency estimates (and all other ratings) for other types of words are less strongly correlated across languages. Therefore, using e.g. English subjective frequency ratings in psycholinguistic studies on Swedish should not be extended to all word types as long as cross-linguistic correlations have not been investigated more thoroughly.

In addition, the results in this paper show that there is a strong correlation between subjective frequency ratings and certain objective lemma frequency measures, in particular frequency counts in blog text corpora. In contrast, there is only a moderate correlation between the subjective frequency estimates and the frequency counts in a so-called balanced corpus (PAROLE). However, as others have argued before me (Gernsbacher 1984, Brysbaert & New 2009, Brysbaert & Cortese 2011), many 'balanced' corpora (at least among those compiled before the 2000s) do not actually represent everyday language use very accurately. Somewhat unexpectedly, the subjective frequency estimates do not correlate any stronger with frequencies in the OpenSubtitles corpus, which has become one of the main sources for objective frequency measures in psycholinguistic studies (thanks to the excellent predictive power of those measures). This indicates that subjective frequency estimates are most likely not very faithful reflections of the objective frequency with which participants encounter words; rather, these estimates may be based on heuristics where several lexical-cognitive variables are confounded. Given that the subjective frequency estimates correlate strongly with word frequencies in children's books, and that children's books are likely to focus more on concrete nouns than what movie dialogues do, one is tempted to speculate that subjective frequency estimates may be affected by variables like concreteness, familiarity and AoA (for a discussion on subjective frequency, familiarity and AoA, see Brysbaert & Cortese 2011). The exact nature of the relationship between these and other psycholinguistic variables, and how they affect the subjective ratings regarding lexical characteristics, requires further concentrated examination by future research.

In comparison with AoA and subjective frequency ratings, the H index for name agreement demonstrates relatively weaker correlations with existing datasets for other languages (.35–.53). In other words, stimuli that elicit relatively uniform responses in Swedish do not necessarily do so in other languages and vice versa. Hence, the choice of appropriate picture stimuli for studies requiring certain levels of name agreement should be based on name agreement norms collected for that specific language. Ideally, to avoid the effects of potential dialectal variation, the name agreement norms should be collected from the very same linguistic population as targeted in the actual experiment.

From a methodological perspective, the current dataset differs in some regards from those collected previously for other languages. First, the naming data in the current dataset was collected using grayscale images in the MultiPic database, while Duñabeitia et al. (2017) used colored images. Nevertheless, the mean of the H index across the 111 items in the current dataset (.57) is very comparable to the corresponding means in e.g. the English (.53), German (.57), and Dutch (.48 and .52 for Netherlands and Belgium, respectively) datasets in Duñabeitia et al. (2017). This suggests that, on average, object recognition is not substantially more difficult from the grayscale images than from the colored images. Second, the data for the current study was collected using an online questionnaire, whereas data acquisition for Duñabeitia et al. (2017) was lab-based, with the exception of the Dutch (NL) dataset. Again, the similarities between the distribution parameters of the H index in the current dataset and those in Duñabeitia et al. (2017) indicate that the lab-based and internet-based data are comparable. Third, and perhaps most significantly, the subjects used in the current dataset are native speakers of Swedish living in a majority Finnish environment (Turku, Finland), while the subjects in Duñabeitia et al. (2017) were speakers of the majority language in the respective region. The Finland-Swedish variety of Swedish contains a considerable amount of Finnish loan words, on the one hand, and a substantial amount of dialectal variation, on the other hand (Ivars 2015). Together, these factors may contribute to some idiosyncrasies regarding the collected norms in comparison with other datasets.

The current dataset will hopefully inspire more psycholinguistic research on Swedish in general and its different varieties, in particular. In addition, this paper hopes to encourage others to collect lexical norms for a wider range of language varieties than what has been the case until now. The current dataset for Finland-Swedish is relatively small both in terms of the number of items and the number of different rating types, and it constitutes, therefore, just an initial step towards facilitating psycholinguistic studies on Finland-Swedish. Future research will do well to expand this dataset by increasing the number and the diversity of the included items (e.g. in terms of semantic categories and concreteness) as well as adding other rating types (e.g. image agreement and arousal).

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S0332586521000123

References

Alario, F.-Xavier & Ludovic Ferrand. 1999. A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers* **31**(3), 531–552.

Ambridge, Ben, Evan Kidd, Caroline F. Rowland & Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language* **42**(2), 239–273.

Baayen, R. Harald, Petar Milin & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiology* **30**(11), 1174–1220.

Balota, David A., Maura Pilotti & Michael J. Cortese. 2001. Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition* **29**(4), 639–647.

Barca, Laura, Cristina Burani & Lisa S. Arduino. 2002. Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, & Computers* **34**(3), 424–434.

Barry, Christopher, Catriona M. Morrison & Andrew W. Ellis. 1997. Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency, and name agreement. *The Quarterly Journal of Experimental Psychology: Section A* **50**(3), 560–585.

Bartolozzi, Federica, Suzanne R. Jongman & Antje S. Meyer. 2021. Concurrent speech planning does not eliminate repetition priming from spoken words: Evidence from linguistic dual-tasking. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **47**(3), 466–480.

Battig, William F. & William E. Montague. 1969. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms. *Journal of Experimental Psychology* **80**(3), 1–46.

Birchenough, Julia M., Robert Davies & Vincent Connelly. 2017. Rated age-of-acquisition norms for over 3,200 German words. *Behavior Research Methods* **49**(2), 484–501.

Bjorklund, David F. & Barbara E. Thompson. 1983. Category typicality effects in children's memory performance: Qualitative and quantitative differences in the processing of category information. *Journal of Experimental Child Psychology* **35**(2), 329–344.

Blomberg, Frida & Carl Öberg. 2015. Swedish and English word ratings of imageability, familiarity and age of acquisition are highly correlated. *Nordic Journal of Linguistics* **38**(3), 351–364.

Boada, Roger, Marc Guasch, Juan Haro, Josep Demestre & Pilar Ferré. 2020. SUBTLEX-CAT: Subtitle word frequencies and contextual diversity for Catalan. *Behavior Research Methods* **52**(1), 360–375.

Bonin, Patrick, Alain Méot, Aurélie Lagarrigue & Sébastien Roux. 2015. Written object naming, spelling to dictation, and immediate copying: Different tasks, different pathways? *The Quarterly Journal of Experimental Psychology* **68**(7), 1268–1294.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp: The corpus infrastructure of Språkbanken. *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC 2012), 474–478.

Boroditsky, Lera, Lauren A. Schmidt & Webb Phillips. 2003. Sex, syntax, and semantics. In Dedre Gentner & Susan Goldin-Meadow (eds.), *Language in Mind: Advances in the Study of Language and Thought*, 61–79. Cambridge, MA: MIT Press.

Boukadi, Mariem, Cirine Zouaidi & Maximiliano A. Wilson. 2016. Norms for name agreement, familiarity, subjective frequency, and imageability for 348 object names in Tunisian Arabic. *Behavior Research Methods* **48**(2), 585–599.

Brysbaert, Marc & Andrew Biemiller. 2017. Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods* **49**(4), 1520–1523.

Brysbaert, Marc & Michael J. Cortese. 2011. Do the effects of familiarity and age of acquisition survive better frequency norms? *Quarterly Journal of Experimental Psychology* **64**, 545–559.

Brysbaert, Marc & Mandy Ghyselinck. 2006. The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition* **13**(7–8), 992–1011.

Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* **41**(4), 977–990.

Brysbaert, Marc, Ilse Van Wijnendaele & Simon De Deyne. 2000. Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica* **104**(2), 215–226.

Cai, Qing & Marc Brysbaert. 2010. SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS One* **5**(6), e10729. DOI: 10.1371/journal.pone.0010729.

Chalard, Marylène, Patrick Bonin, Alain Méot, Bruno Boyer & Michel Fayol. 2003. Objective age-of-acquisition (AoA) norms for a set of 230 object names in French: Relationships with psycholinguistic variables, the English data from Morrison et al. (1997), and naming latencies. *European Journal of Cognitive Psychology* **15**(2), 209–245.

Connine, Cynthia M., John Mullennix, Eve Shernoff & Jennifer Yelen. 1990. Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**(6), 1084–1096.

Cortese, Michael J. & Maya M. Khanna. 2007. Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology* **60**(8), 1072–1082.

Cubelli, Roberto, Daniela Paolieri, Lorella Lotto & Remo Job. 2011. The effect of grammatical gender on object categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **37**(2), 449–460.

Cuetos, Fernando, Andrew W. Ellis & Bernardo Alvarez. 1999. Naming times for the Snodgrass and Vanderwart pictures in Spanish. *Behavior Research Methods, Instruments, & Computers* **31**(4), 650–658.

Dell'acqua, Roberto, Lorella Lotto & Remo Job. 2000. Naming times and standardized norms for the Italian PD/DPSS set of 266 pictures: Direct comparisons with American, English, French, and Spanish published databases. *Behavior Research Methods, Instruments, & Computers* **32**(4), 588–615.

Desrochers, Alain & Glenn L. Thompson. 2009. Subjective frequency and imageability ratings for 3,600 French nouns. *Behavior Research Methods* **41**(2), 546–557.

Dimitropoulou, Maria, Jon A. Duñabeitia, Alberto Avilés, José Corral & Manuel Carreiras. 2010. Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in Psychology* **1**, 218. DOI: 10.3389/fpsyg.2010.00218.

Duñabeitia, Jon A., Davide Crepaldi, Antje S. Meyer, Boris New, Christos Pliatsikas, Eva Smolka & Marc Brysbaert. 2017. MultiPic: A standardized set of 750 drawings with norms for six European languages. *The Quarterly Journal of Experimental Psychology* **71**(4), 808–816.

Elliot, Andrew J. & Markus A. Maier. 2014. Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual Review of Psychology* **65**, 95–120.

Ellis, Andrew W. & Matthew A. Lambon Ralph. 2000. Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**(5), 1103–1123.

Ellis, Andrew W. & Catriona M. Morrison. 1998. Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**(2), 515–523.

Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* **24**(2), 143–188.

Ferrand, Ludovic, Patrick Bonin, Alain Méot, Maria Augustinova, Boris New, Christophe Pallier & Marc Brysbaert. 2008. Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic French words and their relation with other psycholinguistic variables. *Behavior Research Methods* **40**(4), 1049–1054.

Ferrand, Ludovic, Marc Brysbaert, Emmanuel Keuleers, Boris New, Patrick Bonin, Alain Méot, Maria Augustinova & Christophe Pallier. 2011. Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from Chronolex. *Frontiers in Psychology* **2**, 306. DOI: 10.3389/fpsyg.2011.00306.

Gauvin, Hanna S., Magdalena K. Jonen, Jessica Choi, Katie McMahon & Greig I. de Zubicaray. 2018. No lexical competition without priming: Evidence from the picture–word interference paradigm. *Quarterly Journal of Experimental Psychology* **71**(12), 2562–2570.

Gernsbacher, Morton A. 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General* **113**(2), 256–281.

Gilhooly, Ken J. & Robert H. Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation* **12**(4), 395–427.

Glaser, Wilhelm R. 1992. Picture naming. *Cognition* **42**(1–3), 61–105.

Glass, Arnold L. & Keith J. Holyoak. 1974. Alternative conceptions of semantic theory. *Cognition* **3**(4), 313–339.

Greenberg, Michael S. & David F. Bjorklund. 1981. Category typicality in free recall: Effects of feature overlap or differential category encoding? *Journal of Experimental Psychology: Human Learning and Memory* **7**(2), 145–147.

Grigoriev, Andrei & Ivan Oshhepkov. 2013. Objective age of acquisition norms for a set of 286 words in Russian: Relationships with other psycholinguistic variables. *Behavior Research Methods* **45**(4), 1208–1217.

Herdağdelen, Amaç & Marco Marelli. 2017. Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science* **41**(4), 976–995.

Holmes, Selina J. & Andrew W. Ellis. 2006. Age of acquisition and typicality effects in three object processing tasks. *Visual Cognition* **13**(7–8), 884–910.

Ivars, Ann-Marie (ed.). 2015. *Dialekter och småstadsspråk* [Dialects and urban colloquial language] (Svenskan i Finland – i dag och i går 1). Helsingfors: Svenska litteratursällskapet i Finland.

Jescheniak, Jörg D. & Willem J. Levelt. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**(4), 824–843.

Johnson, Carla J., Allan Paivio & James M. Clark. 1996. Cognitive components of picture naming. *Psychological Bulletin* **120**(1), 113–139.

Johnson, Kathy E. 2001. Impact of varying levels of expertise on decisions of category typicality. *Memory & Cognition* **29**(7), 1036–1050.

Johnston, Robert A. & Christopher Barry. 2006. Age of acquisition and lexical processing. *Visual Cognition* **13**(7–8), 789–845.

Juhasz, Barbara J. 2005. Age-of-acquisition effects in word and picture identification. *Psychological Bulletin* **131**(5), 684–712.

Juhasz, Barbara J. & Keith Rayner. 2006. The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition* **13**(7–8), 846–863.

Keller, Dennis & George Kellas. 1978. Typicality as a dimension of encoding. *Journal of Experimental Psychology: Human Learning and Memory* **4**(1), 78–85.

Kiran, Swathi, Katerina Ntourou & Megan Eubank. 2007. The effect of typicality on online category verification of inanimate category exemplars in aphasia. *Aphasiology* **21**(9), 844–866.

Kiran, Swathi & Cynthia K. Thompson. 2003. Effect of typicality on online category verification of animate category exemplars in aphasia. *Brain and Language* **85**(3), 441–450.

Kuhbandner, Christof, Bernhard Spitzer, Stephanie Lichtenfeld & Reinhard Pekrun. 2015. Differential binding of colors to objects in memory: Red and yellow stick better than blue and green. *Frontiers in Psychology* **6**, 231. DOI: 10.3389/fpsyg.2015.00231.

Kuperman, Victor, Hans Stadthagen-Gonzalez & Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* **44**(4), 978–990.

Lambon Ralph, Matthew A. & Sheeba Ehsan. 2006. Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: Empirical and computational evidence. *Visual Cognition* **13**(7–8), 928–948.

Leiner, D. J. 2019. SoSci Survey (Version 3.1. 06). [Computer software] Available at www.soscisurvey.de.

Lété, Bernard & Patrick Bonin. 2013. Does frequency trajectory influence word identification? A cross-task comparison. *Quarterly Journal of Experimental Psychology* **66**(5), 973–1000.

Levelt, Willem J., Ardi Roelofs & Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* **22**(1), 1–38.

Lewis, Michael B. 1999. Age of acquisition in face categorisation: Is there an instance-based account? *Cognition* **71**(1), B23–B39.

Lewis, Michael B., Simon Gerhand & Hadyn D. Ellis. 2001. Re-evaluating age-of-acquisition effects: Are they simply cumulative-frequency effects? *Cognition* **78**(2), 189–205.

Lind, Marianne, Hanne G. Simonsen, Pernille Hansen, Elisabeth Holm & Bjørn-Helge Mevik. 2015. Norwegian Words: A lexical database for clinicians and researchers. *Clinical Linguistics & Phonetics* **29**(4), 276–290.

Lison, Pierre & Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 923–929.

Lotto, Lorella, Luca Surian & Remo Job. 2010. Objective age of acquisition for 223 Italian words: Norms and effects on picture naming speed. *Behavior Research Methods* **42**(1), 126–133.

Łuniewska, Magdalena, Ewa Haman, Sharon Armon-Lotem, Bartłomiej Etenkowski, Frenette Southwood, Darinka Anđelković, Elma Blom, Tessel Boerma, Shula Chiat & Pascale E. de Abreu. 2016. Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods* **48**(3), 1154–1177.

Madden, Daniel L., Martin V. Sale & Gail A. Robinson. 2019. Differentiating beyond name agreement for picture naming: Insight from age-related selection deficits. *Journal of Speech, Language, and Hearing Research* **62**(5), 1373–1380.

Marinelli, Rita, Lisa Biagini, Remo Bindi, Sara Goggi, Monica Monachini, Paola Orsolini, Eugenio Picchi, Sergio Rossi, Nicoletta Calzolari & Antonio Zampolli. 2003. The Italian PAROLE corpus: An overview. *Computational Linguistics in Pisa, XVI–XVII, IEPI., I*, 401–421.

McCloskey, Michael & Sam Glucksberg. 1979. Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology* **11**(1), 1–37.

Medin, Douglas L. & Edward E. Smith. 1984. Concepts and concept formation. *Annual Review of Psychology* **35**(1), 113–138.

Mervis, Carolyn B. & Eleanor Rosch. 1981. Categorization of natural objects. *Annual Review of Psychology* **32**(1), 89–115.

Monaghan, Padraic & Andrew W. Ellis. 2010. Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language* **63**(4), 506–525.

Moors, Agnes, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne & Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods* **45**(1), 169–177.

Morrison, Catriona M., Tameron D. Chappell & Andrew W. Ellis. 1997. 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: Section A* **50**(3), 528–559.

Morrison, Catriona M., Andrew W. Ellis & Philip T. Quinlan. 1992. Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition* **20**(6), 705–714.

Norrby, Catrin, Camilla Wide, Jan Lindström & Jenny Nilsson. 2012. Finland Swedish as a non-dominant variety of Swedish: Extending the scope to pragmatic and interactional aspects. In Rudolf Muhr (ed.), *Non-dominant Varieties of Pluricentric Languages: Getting the Picture*, 49–60. Wien & Frankfurt am Main: Peter Lang Verlag.

Perret, Cyril & Patrick Bonin. 2018. Which variables should be controlled for to investigate picture naming in adults? A Bayesian meta-analysis. *Behavior Research Methods* **51**, 2533–2545.

Pind, Jörgen, Halla Jónsdóttir, Hjördis Gissurardóttir & Frosti Jónsson. 2000. Icelandic norms for the Snodgrass and Vanderwart (1980) pictures: Name and image agreement, familiarity, and age of acquisition. *Scandinavian Journal of Psychology* **41**(1), 41–48.

Postman, Leo & Beverly Conger. 1954. Verbal habits and the visual recognition of words. *Science* **119**(3098), 671–673.

Rips, Lance J., Edward J. Shoben & Edward E. Smith. 1973. Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior* **12**(1), 1–20.

Rofes, Adrià, Lilla Zakariás, Klaudia Ceder, Marianne Lind, Monica B. Johansson, Vânia De Aguiar, Jovana Bjekić, Valantis Fyndanis, Anna Gavarró & Hanne G. Simonsen. 2018. Imageability ratings across languages. *Behavior Research Methods* **50**(3), 1187–1197.

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* **104**(3), 192–233.

Rosch, Eleanor & Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* **7**(4), 573–605.

Rossion, Bruno & Gilles Pourtois. 2004. Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception* **33**(2), 217–236.

Sailor, Kevin M., Molly E. Zimmerman & Amy E. Sanders. 2011. Differential impacts of age of acquisition on letter and semantic fluency in Alzheimer's disease patients and healthy older adults. *Quarterly Journal of Experimental Psychology* **64**(12), 2383–2391.

Samuel, Steven, Geoff Cole & Madeline J. Eacott. 2019. Grammatical gender and linguistic relativity: A systematic review. *Psychonomic Bulletin & Review* **26**(6), 1767–1786.

Schröder, Astrid, Teresa Gemballa, Steffie Ruppin & Isabell Wartenburger. 2012. German norms for semantic typicality, age of acquisition, and concept familiarity. *Behavior Research Methods* **44**(2), 380–394.

Schwanenflugel, Paula J. & Mario Rey. 1986. The relationship between category typicality and concept familiarity: Evidence from Spanish- and English-speaking monolinguals. *Memory & Cognition* **14**(2), 150–163.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423.

Smith, Grinell. 2008. Does gender influence online survey participation? A record-linkage analysis of university faculty online survey response behavior. *ERIC Document Reproduction Service No. ED* 501717 (2008). Available at https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1003&context=elementary_ed_pub.

Snodgrass, Joan G. & Mary Vanderwart. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* **6**(2), 174–215.

Snodgrass, Joan G. & Tanya Yuditsky. 1996. Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments, & Computers* **28**(4), 516–536.

Soares, Ana P., Ana S. Costa, João Machado, Montserrat Comesaña & Helena M. Oliveira. 2017. The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. *Behavior Research Methods* **49**(3), 1065–1081.

Stadthagen-Gonzalez, Hans & Colin J. Davis. 2006. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods* **38**(4), 598–605.

Stanczak, Louise, Gloria Waters & David Caplan. 2006. Typicality-based learning and generalisation in aphasia: Two case studies of anomia treatment. *Aphasiology* **20**(02–04), 374–383.

Tandefelt, Marika (ed.). 2015. *Gruppspråk, samspråk, två språk* [Group varieties, conversations, two languages] (Svenskan i Finland – i dag och i går 2). Helsingfors: Svenska litteratursällskapet i Finland.

Torrance, Mark, Guido Nottbusch, Rui A. Alves, Barbara Arfé, Lucile Chanquoy, Evgeny Chukharev-Hudilainen, Ioannis Dimakos, Raquel Fidalgo, Jukka Hyönä & Ómar I. Jóhannesson. 2018. Timed written picture naming in 14 European languages. *Behavior Research Methods* **50**(2), 744–758.

Tsaparina, Diana, Patrick Bonin & Alain Méot. 2011. Russian norms for name agreement, image agreement for the colorized version of the Snodgrass and Vanderwart pictures and age of acquisition, conceptual familiarity, and imageability scores for modal object names. *Behavior Research Methods* **43**(4), 1085–1099.

Uyeda, Katherine M. & George Mandler. 1980. Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation* **12**(6), 587–595.

Van Overschelde, James P., Katherine A. Rawson & John Dunlosky. 2004. Category norms: An updated and expanded version of the norms. *Journal of Memory and Language* **50**(3), 289–335.

Vitkovitch, Melanie & Lisa Tyrrell. 1995. Sources of disagreement in object naming. *The Quarterly Journal of Experimental Psychology: Section A* **48**(4), 822–848.

Zevin, Jason D. & Mark S. Seidenberg. 2002. Age of acquisition effects in word reading and other tasks. *Journal of Memory and language* **47**(1), 1–29.