

Alex, Alex, or Alex

Do Ambiguous Names Cause Referential Failure Effects?

Bachelor's thesis
presented by

Alexander Clemen

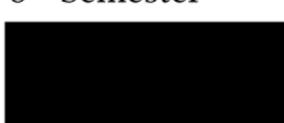
submitted at the
Heinrich-Heine University
Philosophical Faculty
Institute for Linguistics

on the

20th May 2023

Author:

Alexander Clemen 
2832399
8th Semester



alexander.clemen@hhu.de

First Supervisor:

Prof. Dr. Katharina Spalek

Second Reviewer:

Univ.-Prof. Dr. Dr. Peter Indefrey

Advisor:

Prof. Dr. Katharina Spalek

Acknowledgements

I thank Katharina for the many hours of advisory that were not only professionally excellent but also on eye level, making me feel understood and welcome.

I thank my parents for their eternal trust and support.

I thank Jessie for having my back in the most difficult situations.

I thank Anna for all the insightful discussions throughout all stages of this and future works.

Contents

1	Introduction	1
2	Theoretical Background	3
2.1	Anaphora Resolution and Gender in the Mental Lexicon	3
2.2	A Three-Tiered Gender Framework	4
2.3	Previous Findings...	6
2.3.1	... on Gender Cues Using Self-Paced Reading	6
2.3.2	... on Gender Cues in German	8
2.3.3	... Showing Differences in Gender Salience	9
2.4	Stereotypical Gender	9
2.5	Names	10
2.6	Hypotheses	12
3	Methodology	13
4	Norming Study	15
4.1	Method	15
4.1.1	Participants	15
4.1.2	Materials	15
4.1.3	Procedure	17
4.1.4	Selection Criteria	18
4.2	Results and Discussion	18
5	Main and Post Hoc Study	22
5.1	Main Study	22
5.1.1	Participants	22
5.1.2	Materials and Design	22
5.1.3	Procedure	25
5.2	Post Hoc Study	27
5.3	Statistical Analysis	28
5.3.1	Calculating the Main Predictors of Theoretical Interest	28

5.3.2	Data Cleaning	29
5.3.3	Fitting the Best Model	30
5.4	Results	36
5.4.1	Categorical Analysis	36
5.4.2	Continuous Analysis	42
5.4.3	Comparing Significant Regions	43
5.4.4	Split Analyses	43
6	General Discussion	46
6.1	Discussing the Norming Study	46
6.2	Discussing the Main Study	46
6.3	Criticism	49
7	Conclusion	51
A	Materials	v
B	Model Outputs	x
B.1	Summary Outputs of lmers with participant_mm_grouping(_nonAmb) as the Main Predictor of Interest	x
B.2	Summary Outputs of lmers with participant_itemPro_mm_num as the Main Predictor of Interest	xv
B.3	Summary Outputs of the <i>sie</i> -subset Analysis	xx
B.4	Summary Outputs of the <i>er</i> -subset Analysis	xxiv
B.5	Model Diagnostics for the Continuous Analysis	xxvii
C	Glossary	xxix

List of Figures

2.1	Ackerman (2019)'s three-tiered scheme of gender encoding	5
2.2	Semantic system and semantic lexicon of Valentine et al. (1996)'s framework	11
3.1	A flowchart of the three studies, the size of the sample population and the main intended output	14
4.1	Frequency histograms of potentially male, female, and ambiguous names. Names selected for further processing are left of the red line	16
4.2	A dot-plot depicting mean gender rating on the x-axis by subjective gender rating on the y-axis for each participant ($n = 35$). The colour represents the potential gender which was determined by the Google queries	19
4.3	Mean gender rating split by potentially male (red), ambiguous (green), and female (blue) names. The rectangles describe a selection criterium and the grey shade shows all data in all facets	20
4.4	A comparison of the distribution of mean gender ratings for role names or proper names by their standard deviation	21
5.1	The distribution of filler items and their mean gender rating taken from Ken-nison and Trofe (2003) vs the optimal distribution of mean gender rating	23
5.2	Procedure of the self-paced-reading experiment. The comprehension ques-tions were shown 25% of the time	27
5.3	Correlation matrix of all potential predictor variables	31
5.4	Normal probability plot for the final model at region 04	35
5.5	Q-Q plot and residual plot for the final model at region 04	35
5.6	Linearity and Homoscedasticity plot for the final model at region 04	36
5.7	Reading time for each condition of <code>participant_mm_grouping(_nonAmb)</code> for presentation regions 01 to 07	41
5.8	Distribution of reading time at region 04 and 05 with <code>participant_mm_grouping</code> (Match, Mismatch, Ambiguous) as the main predictor of interest and reading time increase dependent on the increase in Mismach value <code>partic-pant_itemPro_mm_num</code>	44
B.1	Model diagnostics in the Continuous Analysis	xxviii

List of Tables

2.1	Table of Irmens and colleagues' stimuli (levels) and an example	8
5.1	An example sentence, its chunks, the regions of interest, and the region (reg) number. "Pronoun" is shortened to "Pro" and $x+n$ indicates how many regions (n) the present region is presented after x	24
5.2	Item-Pronoun combination in carrier sentences throughout six lists; M = Stereotypical Male Item, F = Stereotypical Female Item, A = Ambiguous Item, m = masculine Pronoun (dark colour), f = feminine Pronoun (light colour)	24
5.3	An example of comprehension questions for each targeted region (reg)	25
5.4	Potential results from Post Hoc study and their incorporation with the Main study results evaluation	29
5.5	List of potential predictor variables, their type, and values. Crossed-out variables were not included in the best model	30
5.6	AIC values for lmers with different data frames at the most important region .	32
5.7	AIC comparisons between full model, and models with fewer predictor variables	33
5.8	AIC comparisons between full model and model without item as random effect	33
5.9	Summary outputs of the Linear Mixed-Effects Models for regions 01 and 02 .	38
5.10	Summary outputs of the Linear Mixed-Effects Models for regions 04 and 05 .	39
5.11	Summary outputs of the Linear Mixed-Effects Models for regions 06 and 07 .	40
5.12	Table of coefficients of determination for all regions of interest	42
5.13	Summary outputs of the <i>er</i> -subset and <i>sie</i> -subset analyses	45
A.1	List of rated names in ascending mean gender rating order	viii
A.2	List of filler items (role names translated from English (Kennison & Trofe, 2003)) in ascending mean gender rating order	ix

Chapter 1

Introduction

Lexical and syntactic ambiguity have been central for many decades of psycholinguistic research, but to the best of my knowledge, not one paper questioned the use of ambiguous first names. Consider the sentence *Alex liest diese Thesis. Er denkt, dass das Thema relevant ist.* (*Alex is reading this thesis. He thinks the topic is relevant.*) one does not know if *er* is the “right” pronoun for *Alex* because *Alex* could be stereotypically male, female, or both. In research on anaphora resolution, gender cues have helped to better understand coherence and coreference (Kehler et al., 2008), accessibility (McKoon et al., 1993), prominence (Swaab et al., 2004), implicit causality (Garnham et al., 1992), and syntactic constraints (Sturt, 2003) but some stimuli used in these papers might be intrinsically flawed. Names such as *Tony, Sam, Freddy, Max*, or my own name *Alex* – the short form of *Alexander* and *Alexandra* – seem inherently ambiguous, such that gender cues might malfunction. If *Alex* entails two stereotypical genders – male and female – the pronouns *er* (*he*) and *sie* (*she*) could always mismatch in gender because one gender is always violated, or the anaphora and the antecedent always match because one of *Alex*’s genders is always match with the gender cue from the pronoun. This gap in research and the personal connection of my own name motivated me to the research question:

Do ambiguous first names cause referential failure effects?

This thesis connects three linguistic topics: gender, names, and anaphora resolution. In linguistic theory, **gender** has been incorporated in concept nodes at the lexical level (Levelt et al., 1999) and feature nodes at the feature level (Dell & O’Seaghdha, 1992). The Ackerman (2019) framework presents an approach that uses exemplar and prototype theory (Medin & Schaffer, 1978; Rosch & Mervis, 1975). Gender can be abstracted from a continuous exemplar tier, but it can also be interpreted in categories from a category tier. In this work, I will investigate gender, like Ackerman (2019), as a category of stereotypical male, female and ambiguous names, but also consider gender on a continuum. On the study of **names**, Valentine et al. (1996) have presented evidence that names are in some respects similar to words and in other respects similar to faces. I will provide evidence that names differ from role names –

words that describe a person’s occupation (e.g. *accountant*), a field of general activity (e.g. *student*), or other means of description (e.g. *wife*). Previous work on **anaphora resolution** has utilized self-paced reading (Cacciari et al., 1997; Carreiras et al., 1993; Irmén & Kurovskaja, 2010; Kennison & Trofe, 2003), Eye-Tracking (Irmén & Schumann, 2011), and EEG (Hammer et al., 2005; Irmén et al., 2010; Osterhout, 1997; Schmitt et al., 2002) and showed longer reading time, longer fixation time, and informative Event-Related Potentials if the anaphora and antecedent misaligned in gender.

This thesis is modelled after Kennison and Trofe (2003)’s self-paced reading paradigm, which investigated gender mismatch effects for role names and personal pronouns. I use male, female and ambiguous first names instead and expect to find differences in reading time between the Match, Mismatch, and Ambiguous conditions which express different degrees of name–pronoun gender (mis)alignment.

In the following section, Chapter 2, I will give a broad overview of anaphora resolution theories. Then I will examine the Ackerman (2019) framework in more detail and present the aforementioned experimental research on anaphora resolution. This will allow me to argue for a broad definition of stereotypical gender. Research showing the special status of names is presented after that. I will derive my hypotheses for the subsequent experiments from this comprehensive overview.

Chapters 3 to 5 lay out the three experiments I conducted following Kennison and Trofe (2003). A Norming study was conducted since no normed material existed. I created a name corpus of over 11,000 tokens and had participants rate the “best” 143 names to create an objective name–gender association account. The Main study consisted of a self-paced reading experiment measuring referential failure effects. Stereotypically male, female, and ambiguous names were crossed with grammatically masculine and feminine personal pronouns. An extension of Kennison and Trofe (2003)’s study design is the Post Hoc study, which captured subjective name–gender ratings used for later statistical analysis. The rating allows me to evaluate gender as a category but also as a continuum.

In Chapter 6, I will discuss the findings from the Norming and Main study in the light of previous work and discuss their statistical and methodological limitations. In Chapter 7 I will draw a conclusion, make a recommendation to handle gender as a variable, and close with a perspective on future research.

Chapter 2

Theoretical Background

In this chapter, I will give an overview of theories and findings relevant to this thesis. First, I will present theories on anaphora resolution and a framework for gender encoding. Then, I will present previous findings on anaphora resolution with the same method I use, findings with the same language I use, and findings which illustrate another layer of complexity for the use of gender in research. From this, I will draw my definition of gender in this thesis. Finally, I will present a model of name recognition and reasons why names are a special category of words.

2.1 Anaphora Resolution and Gender in the Mental Lexicon

Accessibility theory, as presented by Ariel (1991), describes anaphoras as accessibility markers that match their antecedents' degree of accessibility in our memory. The degree of accessibility (acc.) of discourse entities is mediated by three criteria:

- **Informativity** is the degree of lexical and semantic information in a marker. While “Pass it.” (high acc.) points to one salient entity, “Pass the yellow book.” (low acc.) points to one entity of potentially many competing options.
- **Rigidity** is the degree of how uniquely identifiable an entity is. While “He is a president.” (high acc.) points to all male individuals, “Donald Trump is a president.” (low acc.) points to one individual without any context needed.
- **Attenuation** describes the degree of how detailed or focused information is presented. “The United Kingdom is no EU country.” (high acc.) puts more stress on the country than “The UK is no EU country.” (low acc.) and a *stressed it* (low acc.) and an *unstressed it* (high acc.) point do different degrees of accessibility.

For entity retrieval, accessibility markers are set to a specific degree of accessibility following a hierarchy, laid out in Ariel (1991: 449), which is language dependent¹.

¹For example, zeros are possible in Spanish but not in German or English.

A second account for determining the reference of anaphoras is presented by Gordon et al. (1993). The researchers provide a framework that tries to explain local discourse and the importance of pronouns in discourse. Centering theory, as presented in Grosz et al. (1986), has at its core two principles. (i) Discourse entities are forward-looking centres (Cf) and are ranked in terms of their prominence. The most prominent Cf is the backward-looking centre (Cb) and is the centre of the discourse. These centres (i. e. anaphoras) indicate that one utterance is coherent with the previous utterance and are used top-down in order of prominence. (ii) If the second-highest Cf is realized as a pronoun, then the highest Cf (i. e. the Cb) must be realised as a pronoun too, because the Cb is most prominent, and pronouns are used starting with the highest prominence. Gordon et al. (1993) have shown that there is a loss in reading time (repeated-name penalty) when the Cb is referred to twice with a name instead of first as a name and second as a pronoun. The repeated-name penalty illustrates that pronouns establish coherence between sentences (Gordon et al., 1993).

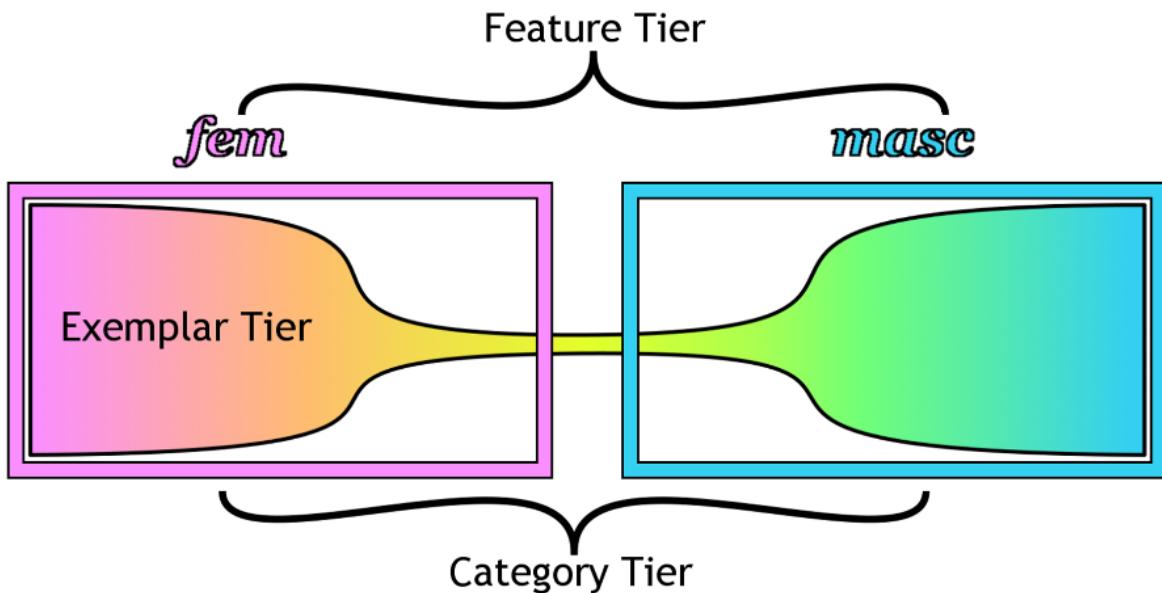
The two theories seem to agree that anaphoras are referring expressions with little meaning of their own but rather pointing devices to either a degree of accessibility or prominence. Anaphoras maintain discourse integrity with the help of featural clues (e. g. gender or number agreement), but even if anaphoras lack phonetic content completely (e. g. zeros) – and as such, all featural clues – the absence acts as a pointing device (Callahan, 2008: 239). In this thesis, I only use pronouns as anaphoras and use *pronoun* and *anaphora* interchangeably.

2.2 A Three-Tiered Gender Framework

When a pronoun agrees in gender with its antecedent, how was the gender encoded in the antecedent in the first place? A framework that covers many aspects of gender acquisition, gender classification, and the use of grammatical gender dependent on culture and language was first presented by Ackerman (2019: 116) in a three-layered scheme (see Figure 2.1).

The exemplar tier consists of an individual's observations. These observations consist of tokens of gender expressions as a matrix of property (e. g. hair, voice) and expression (e. g. length, pitch). Individuals who grew up in a binary gender expression environment have received a gender binary set of observations, and individuals who grew up in non-conforming communities have gender cues distributed differently in the exemplar tier. In Figure 2.1, the exemplar tier is the double-vase-shaped continuum and depicts that many properties and their expressions fall onto the outermost spectrum describing maleness/femaleness, and properties that do not align with these binary expressions are found on the handle on the vase-shaped continuum.

Figure 2.1: Ackerman (2019)'s three-tiered scheme of gender encoding



The **category tier** and their gender categories are established bottom-up from the input from the exemplar tier (e.g. own observation) and top-down through semantics²). As such, depending on the input – which relies on culture and experience – there may be two distinct gender categories (<MALE> and <FEMALE>) or more, but essentially they never overlap. This explains why androgynous individuals are difficult to classify. A shift of category boundaries is slow since every new piece of information makes up an increasingly smaller fraction of the whole such that early childhood input is of the greatest importance for gender category development.

Ackerman (2019) presents three options for how lexical gender assignments are formed: (i) An interaction of the exemplar tier and the category tier, which generates a probability of genderedness for a lexical item, (ii) the existence of an aggregate of lexical items coupled each with a specific gender category³, (iii) or the lexical item and gender are associated in the mental lexicon at some point with rare updating. Lastly, the author proposes conceptual gender mapping does not necessarily need to be a one-to-one mapping (*Phillip*-<MALE> or *Anna*-<FEMALE>) but can be a one-to-many mapping (*Alex*-<MALE/FEMALE>) and the conceptual gender information is stored separately from the grammatical gender information.

The **feature tier** stores the grammatical gender as a pure linguistic marker not strictly related to the conceptual gender. For German, there are three noun classes, whereas, for most role names and proper names, *neuter* (N) is ungrammatical because usually *masculine* (M) and

²Gender schema refers to a cognitive framework that individuals use to organize and process information related to gender. The schemas (e.g. “Men are aggressive and women do the dishes.”) are learned through socialization processes, such as observing and imitating gender roles and behaviours, and are reinforced by social institutions, such as family, media, and education. I am aware that the influences of bottom-up inputs and top-down inputs overlap, but I took the example of gender schema from the author.

³At comprehension a lexical item is drawn at random.

feminine (F) antecedents are human referents.

Ackerman (2019) proposes that all three tiers are viable for co-reference, but languages which make little use of grammatical gender, such as English, use the category tier for co-reference while German, a language which has strict gender agreement, use the feature tier during co-reference. Previous research will show if this assumption is true.

2.3 Previous Findings...

2.3.1 ... on Gender Cues Using Self-Paced Reading

Various studies investigated the effect of stereotypical gender agreement violations (e. g. *surgeon_{<MALE>}* ... *she_F*) using self-paced reading, Eye-Tracking and EEG. For a brief overview of self-paced reading experiments that used personal pronouns referring to one animate antecedent, I will summarize Cacciari et al. (1997), Carreiras et al. (1996), and Kennison and Trofe (2003). All three papers introduced an entity with a role name in the first sentence which was referred to in the second or third sentence with a gender-matching or mismatching personal pronoun.

Carreiras et al. (1996) investigated stereotypically male, female and neutral role names followed by grammatically masculine or feminine co-referring personal pronouns in English (see Example (1-a)) and Spanish (see Example (1-b)). In their first experiment (in English) they found longer mean reading times when the grammatical gender of the anaphora mismatched the stereotypical gender of its antecedent. The mean reading time of sentences with pronouns referring to neutral role names (e. g. *the student_{<Ø>}* ... *he_M/she_F*) was no different from the Match condition (p. 645).

In Spanish, disambiguation can already be resolved within the role name NP. The stereotypical gender of the role name matched or mismatched with its preceding grammatically gendered article (see Example (1-b)). The first sentence was read slower in the Mismatch condition than in the Match condition and the second sentence, containing the pronoun⁴, was read equally fast independent of the role names stereotypical gender. The Neutral condition (e. g. *El/La abogado/a (The lawyer)*) equalled the Match condition in both sentences (p. 650). The authors conclude gender information is activated at the earliest possible point and that the grammatical gender of the article overwrites the stereotypical gender from the role name in the situation model of the reader. Consequently, anaphora resolution processes do not suffer even if the character's stereotypical gender mismatches the pronoun's grammatical gender. The authors believe that most role name tokens incorporate stereotypical gender information and if a role name has no stereotypical gender information, the assignment of gender is as fast as a matching gender update (p 657).

⁴The grammatical gender of the pronoun was always identical to the grammatical gender of the article.

- (1) a. *The electrician_{<MALE>} examined the light.*
He_M/She_F needs a special attachment to fix it.
- b. *El_M/La_F carpintero/a_{<MALE>} tomó las medidas para hacer el armario. [...] (The carpenter took measurements to make the cupboard.)*
El_M/Ella_F tenía que terminarlo en el plazo de una semana. (He/She had to finish in the space of one week.)

Cacciari et al. (1997) investigated functionally ambiguous Italian words. “Epicens” (see Example (2-a)) take one grammatical gender denoted in the article but can conceptually refer to both males and females and “ungendered words” (see Example (2-b)) are grammatically and conceptually gender opaque⁵. For “epicens” the grammatical gender could match or mismatch while “ungendered words” always matched. Both types of words always matched stereotypical gender. Reading times (amongst “epicens”) were faster when the pronoun matched the antecedent’s grammatical gender than when it did not. Matching “epicens” were also faster read than the always matching “ungendered words”⁶. The authors conclude that even though conceptual gender always matches for ambiguous words, there is a facilitatory effect when the grammatical gender matches but no reading time penalty when the anaphora grammatically mismatch its antecedent.

- (2) a. *La_F vittima_{<MALE/FEMALE>} dell’incidente stradale sbatté violentemente la testa contro il finestrino. (The victim of the car accident violently slammed the head against the window.)*
Lei_F/Lui_M, perciò, perse molto sangue e svenne. (She/He, therefore, lost a lot of blood and fainted.)
- b. *L’_{M/F}erede_{<MALE/FEMALE>} decise di andare in vacanza con i soldi ricevuti dalla zia. (The heir decided to go on vacation with the money.)*
Lei_F/Lui_M, perciò, progettò un lungo viaggio negli USA. (She/He, therefore, planned a long trip to the States.)

Kennison and Trofe (2003) conducted a comprehensive rating study (405 role names), and a self-paced reading study with phrase-by-phrase presentation moving windows⁷, which allowed for a more fine-grained analysis than Carreiras et al. (1993). Kennison and Trofe (2003) found in the first and second regions after the pronoun that the mean reading time was longer in the Mismatch condition compared to the Match condition. The mean reading time was not significantly different at the pronoun region. Further, *she* was read on average slower than

⁵They are gender opaque due to a reduced article (*l'*) and the absence of a morphological gender marker (-a or -o).

⁶Note that Cacciari et al. (1997: 523) describes *t(22), p = .10* as “close to significant.”

⁷Moving window means that the whole sentence was shown as a dotted line except for the presentation region, which was shown as text. When the participant moved to the next region, the previous region turned back into the dotted line.

he (Kennison & Trofe, 2003: 364), which is explained by the printed frequency of the word, and effects for stereotypical gender were found for sentence-final words, explained by the sentence wrap-up effect (Just & Carpenter, 1980: 331). The researchers did not investigate gender-neutral stimuli. They conclude that speakers have a stereotypical gender representation mapped to every word in the mental lexicon.

- (3) *The executive_{<MALE>} *distributed *an urgent *memo.*
*He_M/She_F * made it clear *that *work *would continue *as normal.**

Note: The asterisk indicates the region boundary.

2.3.2 ... on Gender Cues in German

Lisa Irmens and colleagues conducted a series of experiments using reading time and fixation time to understand gender cues in discourse better. The type of anaphora and antecedent with corresponding stereotypical genders are displayed in Table 2.1.

Table 2.1: Table of Irmens and colleagues' stimuli (levels) and an example

Paper	Antecedent	Anaphora
Irmens and Kurovskaja, 2010	role name (male, female, neutral)	role name (male, female)
	<i>Dieser Kassierer_{M<FEMALE>} ist mein Mann_{M<MALE>}.</i> (<i>This cashier is my husband.</i>)	
Irmens and Schumann, 2011	role name (male, female)	role name (male, female, neutral)
	<i>Mein Bruder_{M<MALE>} ist Sänger_{M<Ø>} in einer Band.</i> (<i>My brother is a singer in a band.</i>)	

In a self-paced reading experiment Irmens and Kurovskaja (2010: 372) found the same effects as Carreiras et al. (1996) and Kennison and Trofe (2003) even though the referring expressions were a role names and not personal pronouns. The results of gender-neutral antecedents were not discussed.

Using Eye-Tracking, Irmens and Schumann (2011) showed longer fixation times when there was an incongruity between the antecedent's and the anaphora's stereotypical gender compared to the Match condition. A novel finding was that female role names had immediate resolution while male role names showed late resolution. The researchers believe that singular masculine role names (e.g. *Sänger*) are gender ambiguous (or underspecified) and singular female role names (e.g. *Sängerin*) are gender unambiguous. Immediate resolution for unambiguous role names is in line with previous findings (Duffy & Keir, 2004; Sturt, 2003).

2.3.3 ... Showing Differences in Gender Salience

On the difference between grammatical gender and stereotypical gender Schmitt et al. (2002) made use of the fact that all German diminutives are grammatically neuter. Diminutives (e. g. *Bübchen_{N<MALE>}* (*little boy*)) and non-diminutives (e. g. *Bub_{M<MALE>}* (*boy*)) were crossed with the German three personal pronouns (*er_M* (*he*), *sie_F* (*she*), *es_N* (*it*)) allowing for anaphora resolution in which both stereotypical gender and grammatical gender matched or mismatched (“double violation”). This design also allowed that only one gender did not match (“single violation”). For double violations with *er/sie* referring to non-diminutives, the researchers found Event-Related Potentials (ERPs) indicating semantic processing and syntactic reanalysis. Double violations with the a-typical *es* referring to non-diminutives showed that syntactic processes but no semantic processing were involved. Diminutives, on the other hand, showed ERPs indicating syntactic reanalysis no matter if stereotypical gender was violated, grammatical gender was violated, or both were violated. Based on the results, Schmitt et al. (2002) conclude that the “biological gender” is less salient for diminutives than for non-diminutives since it is relevant only for non-diminutives.

Similar effects were found in Hammer et al. (2005) and Osterhout (1997). Hammer et al. (2005) compared animate and inanimate antecedents (in German), in which the effect, indicating syntactic reintroduction, was stronger for animate than inanimate antecedents. Osterhout (1997) found (in English) a larger ERP amplitude of gender mismatching pronouns when they referred to role names with “definitional gender” (e. g. *father, mother*) than when they referred to role names with stereotypical gender (e. g. *surgeon, nurse*).

Bjorkman (2017) discussed the use of the gender-neutral singular *they* and indirectly presents a scale of acceptability (see Scale (4)). Collections of people and visually or auditory unidentifiable individuals are generally acceptable referents for *they*; second, indefinite role names and kinship nouns with underspecified gender are for some English speakers acceptable, while for others not (indicated by “%”); third, definite role names and ambiguous proper names are acceptable by “[s]ome innovative *they* users” (Bjorkman, 2017: 6); fourth, the use of *they* with kinship nouns with definite gender and unambiguous proper names is ungrammatical in English.

- (4) everyone; Jannet and Tomas; unidentified person <%the professor; %child/cousin <%Prof. Smith; %Alex/Chris; %moongirl17 < *sister/father; *Jannet/Tomas (Bjorkman, 2017)

2.4 Stereotypical Gender

Much of the aforementioned works point to the idea that gender is more complex than match or mismatch. Kennison and Trofe (2003: 366) defines stereotypical gender as “the relative frequencies [a role name] coincides with the association male or female.”, which seems to be

in line with the Ackerman (2019) framework. But why should the term stereotypical gender be limited to role names? A *father* with “definitional gender” male could just simply be more stereotypically male than a *surgeon* so the relative frequency of the co-occurrence of *father* and male was higher than *surgeon* and male (explaining the difference in Osterhout (1997)) while the “gender neutral” *cousin* simply had a less frequent co-occurrence with the concept of male. Indeed referring expressions such as *Anna*, *Phillip* and *Alex* are also only used to refer to males, females or non-binary people and could fit under the umbrella term “stereotypical gender”. *Anna* or *Phillip* would be stereotypically unambiguous like *father* or *mother* while ambiguous names like *Alex* would be stereotypically ambiguous like *cousin*.

For this thesis, I will use “stereotypical gender” with the definition proposed by Kennison and Trofe (2003) with the addition that there is a degree of gender salience (cf. Hammer et al. (2005), Osterhout (1997), and Schmitt et al. (2002)) for every word in the lexicon. “Stereotypical gender” works as an umbrella term for “semantic gender” (Irmens & Schumann, 2011), “conceptual gender” (Ackerman, 2019), “definitional” or “lexical gender” (Cao & Daumé, 2021; Kreiner et al., 2013) which all associate the lexical items to male (<MALE>) or female (<FEMALE>) properties.

2.5 Names

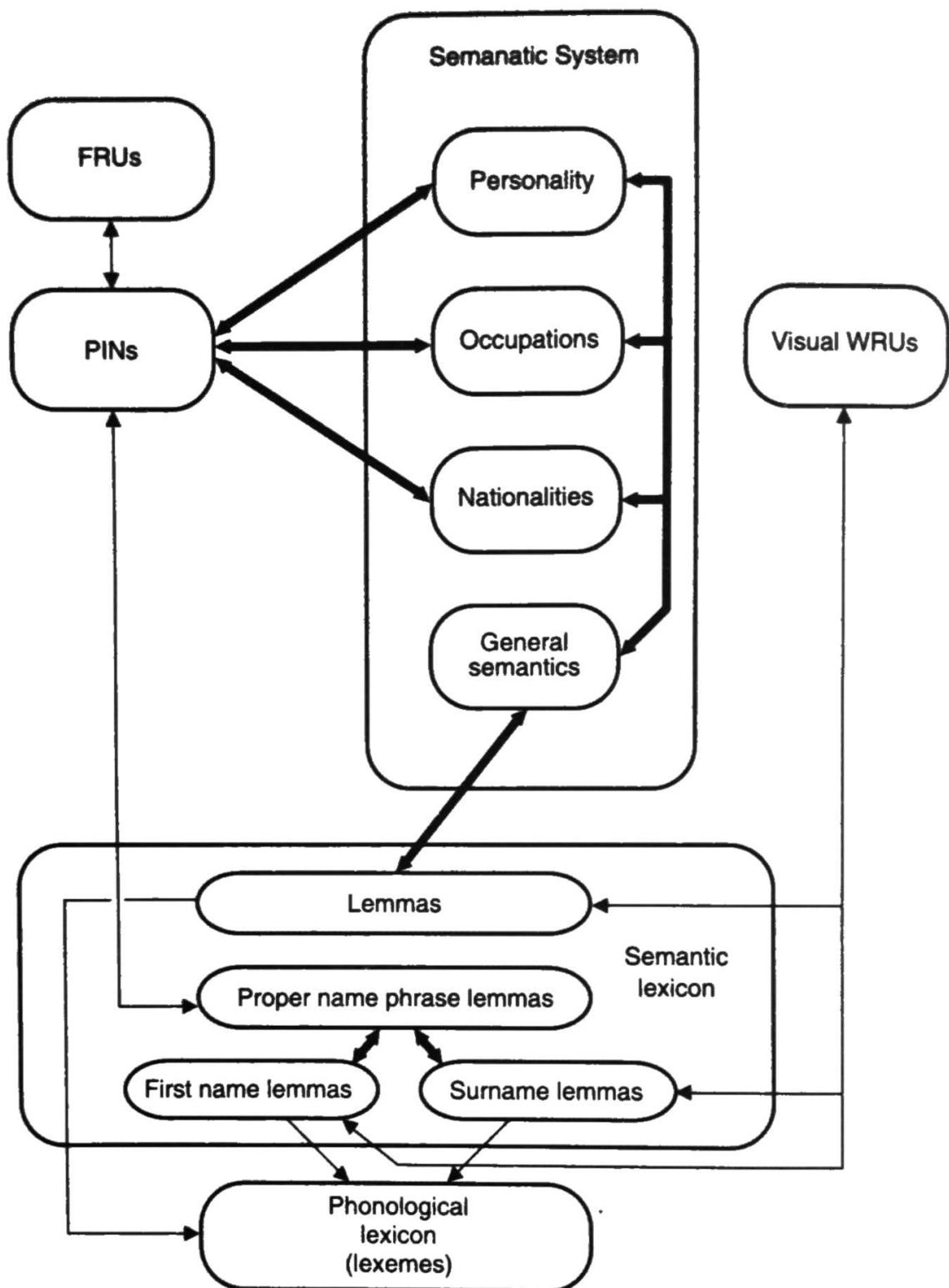
For this thesis, people’s first names (or proper names and names) are “pure referring expressions” (Semenza & Zettin, 1988) that have *reference* but no *sense* (Frege, 1948) and thus are relatively meaningless (Cohen & Burke, 1993).

Valentine et al. (1996: 172) present a framework that illustrates the meaninglessness but also the unique status of names. Figure 2.2 shows that a person’s name (Proper name phrase lemma) is separate from other lemmas in the semantic lexicon, and both, a First name lemma or a Last name lemma, can activate it⁸. A Proper name phrase lemma is not directly connected to its semantics but first needs to activate its Person Identity Node (PIN), which is a token that connects to the identity-specific semantics in the Semantic System and, as such, the features ascribed to that person.

The separation of name lemmas and other lemmas explains why the tip-of-the-tongue (TOT) phenomenon is significantly more frequent for names than for other words (see Cohen and Burke (1993: 251) for a summary). The differentiation between the two is underpinned by Semenza and Zettin (1988), who describe a patient with the inability to report proper names while other word naming abilities stayed intact. Lastly, Valentine et al. (1991: 173) showed that dependent on task, name recognition shares aspects of face recognition (distinctiveness effects) and word recognition (frequency effects), differentiating names from other words.

⁸The first name or last name lemmas are activated through auditory or visual Word Recognition Units (WRUs).

Figure 2.2: Semantic system and semantic lexicon of Valentine et al. (1996)'s framework



2.6 Hypotheses

The previous findings speak for a scale of mismatch. Both *father...she* and *surgen...she* showed gender violation effects, but *father* caused a stronger amplitude than *surgeon*. Role nouns that are lower in the mismatch scale (e. g. *student...she*) showed no effect. Due to the ubiquity of names in our day-to-day life and the high rigidity of names (Ariel, 1991), I expect a general tendency of first names causing greater effects than role names. Consequently, there should be a significant effect for gender violations of unambiguous as well as ambiguous names, but the effect of ambiguous names should be less strong.

I am measuring gender violation effects in reading time on a word-by-word basis; hence the two hypotheses formulated below have each four sub-hypothesis for every word/region that is measured. Previous studies (Chow et al., 2014) have shown that gender violation effects can have spillover effects of up to two words/regions. Due to the importance of names, I have formulated my hypothesis for up to three spillover words/regions.

H1: [a: pronoun; b: first spillover; c: second spillover; d: third spillover]

The mean reading time at the [a/ b/ c/ d] region is significantly longer in the Mismatch condition than in the Match condition.

H2: [a: pronoun; b: first spillover; c: second spillover; d: third spillover]

The mean reading time at the [a/ b/ c/ d] region is significantly longer in the Ambiguous condition than in the Match condition.

Research suggests that there is a strong tendency to assign an initial default gender (Cacciari et al., 1997: 518) but for ambiguous names, this means that multiple Person Identity Nodes are activated, which activate conflicting gender information. This should increase cognitive load more than for single-gendered referring expressions. Hence I have formulated the following hypothesis for the regions in which the name and the subsequent word are presented.

H3: [a: item; b: item spillover]

The mean reading time in the [a/ b] region is significantly longer in the Ambiguous Condition than in the Non-Ambiguous Condition.

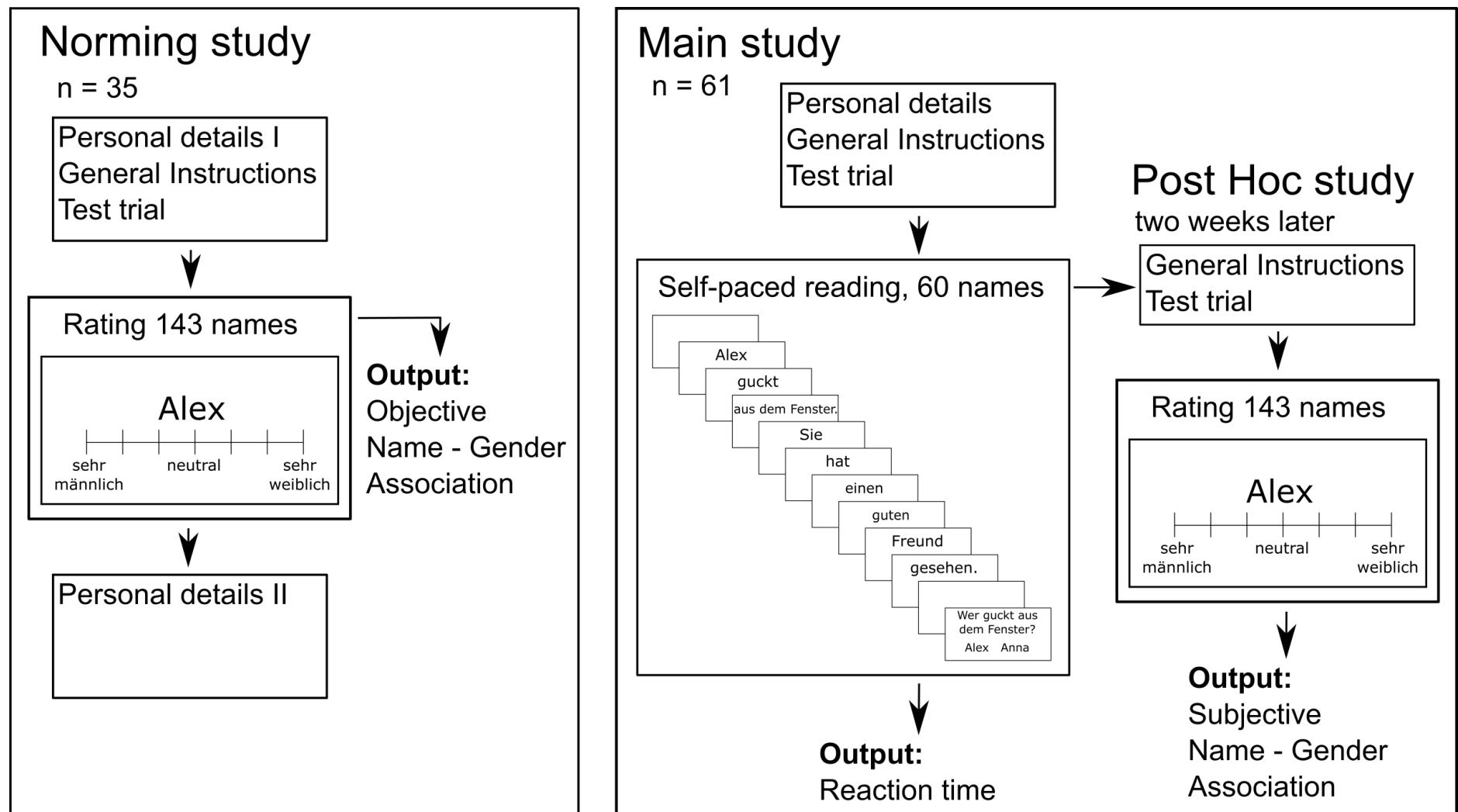
Chapter 3

Methodology

This thesis is largely modelled after Carreiras et al. (1996) and Kennison and Trofe (2003)'s experimental setup, and Figure 3.1 illustrates the sequence of the three studies I conducted. Due to a lack of publicly available normed material, I conducted a Norming study (see Chapter 4) and used the objectively best male, female and ambiguous names in the Main study. I conducted a self-paced reading experiment in the style of Kennison and Trofe (2003) and measured potential reading time differences during gender encoding and anaphora resolution (see Section 5.1). The Post Hoc study was analogous to the Norming study in its experimental design and captured the participant-specific name–gender association (see Section 5.2). The name–gender association allowed me to capture and calculate key components. On the one hand, I could gain information on the confounding variables “attitude towards gender” and “gender of friends and family”, which I could not include in the demographic form before the Main study experiment. On the other hand, the participant-specific variable allowed me to calculate the Match, Mismatch, and Ambiguous conditions for the statistical analysis (see Section 5.3).

Detailed descriptions of the quantitative research, including data collection, data transformation, and the results yielded, are found in the sections of the experiments. The participants of this study did not give written consent for their data to be shared publicly. Due to the sensitive nature of the research, supporting data is not available, but the code and experimental data are available online (see Section 7), and many lists and tables are found in Appendix A.

Figure 3.1: A flowchart of the three studies, the size of the sample population and the main intended output



Chapter 4

Norming Study

Most research on anaphora resolution avoids discussing the gender of first names as antecedents. Researchers either use role names (Kennison & Trofe, 2003), decide that a name is male or female (Nieuwland et al., 2007: 995), or use first names of celebrities (Nieuwland et al., 2007: 995). Due to these workarounds, there is, to the best of my knowledge, no database of first names and a corresponding gender rating. But this rating, as an objective measure to select items, is essential for my Main study; therefore, I conducted a norming study. First, I will describe how I compiled a corpus and selected potentially male, female, or ambiguous names. Second, I will describe how I conducted the rating study, selected the objectively best-rated 60 names, and kept the rest for other purposes¹. Lastly, I will discuss the advantages of the Norming study compared to the direct usage of web queries for male, female, or ambiguous names.

4.1 Method

4.1.1 Participants

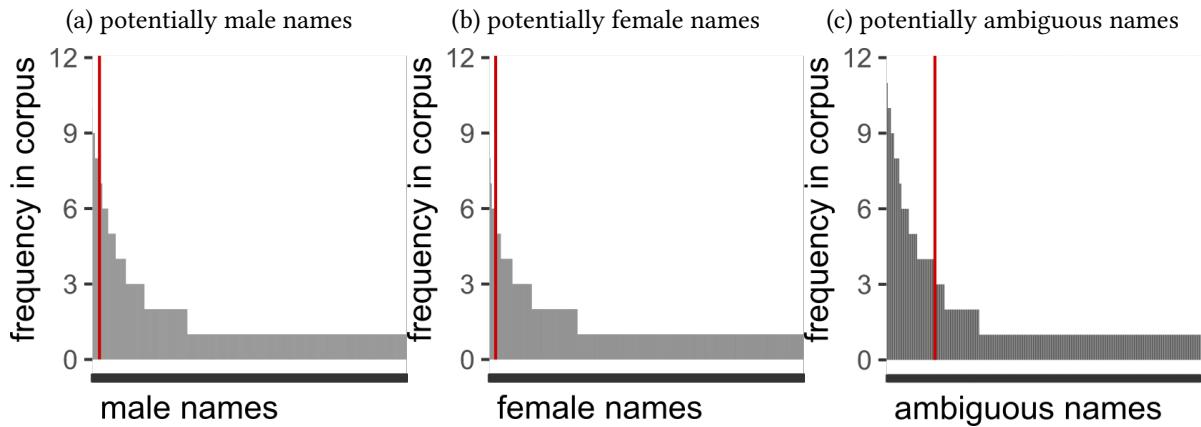
35 (i) German native speakers, (ii) between the age of 18 and 35 (30 females, 4 males, and 1 non-binary person; mean age: 25.12 years (*range*: 17 – 30 years, *sd*: 3.96 years)) from the university rated 143 names (visualized in Figure 3.1) on a 7-point rating scale.

4.1.2 Materials

To minimize the designer's bias, I created a corpus of names based on the most relevant websites for people's first names. I compiled a list of all names on the first ten websites found on Google for the search terms “unisex Namen” (*unisex names*), “geschlechtsneutrale Namen” (*gender neutral names*), “Namen für Mädchen” (*names for girls*) and “Namen für Jungen” (*names for boys*). Since “unisex Namen” and “geschlechtsneutrale Namen” suggested

¹I used 83 names in warm-up trials and alternative answers in comprehension questions.

Figure 4.1: Frequency histograms of potentially male, female, and ambiguous names. Names selected for further processing are left of the red line



baby names I searched for “Namen für Mädchen” instead of “Namen für Frauen” (*names for women*) and “Namen für Jungen” instead of “Namen für Männer” (*names for men*). This ensured that there was no inherent age difference between potentially ambiguous names and potentially unambiguous names. The final corpus consisted of 11.208 tokens.

I grouped names with minor orthographic differences as one (e. g. [*Philip*, *Phillip*, *Philipp*] → *Philipp*²). Afterwards, I performed a frequency analysis using R (Version 4.1.2; R Core Team (2022)) and selected the most frequent male, female, and ambiguous names and tried to match best the group size (60 potentially male names, 57 potentially female names, 45 potentially ambiguous names) (see Figure 4.1). To further match the group size and reduce the set of 162 names, I set three selection criteria:

- The name *Alex*, *Alexander*, *Alexandra*, or variant forms were not put into the final set of names because *Alexander* is the researcher’s name, and the participants will have communicated with me, which potentially primed the name male.
- If there was a short form and a long form of a name (e. g. *Mats* vs *Matthias*), the form with the higher frequency was preferred, unless ...
- ... the comparison was with a potentially ambiguous name. Then the potentially ambiguous name was preferred since there were fewer ambiguous names in the data frame.

The final list of names after the first selection process consisted of:

- 52 potentially male names
- 41 potentially ambiguous names
- 50 potentially female names

²The most frequent version of the name was used, such that *Philip* and *Phillip* were changed to *Philipp*.

4.1.3 Procedure

All studies were created using *PsychoPy3* (Peirce et al., 2019) and hosted by and conducted with *Pavlovia*. A welcoming screen greeted the participants, and the participant's first name and age were inquired about on subsequent screens. Afterwards, two screens stated the goal and the procedure of the study:

(Introduction screen 1)

Ich versuche zu verstehen wie Vornamen im mentalen Lexikon abgespeichert sind und dafür brauche ich (d)eine Datengrundlage. Vorerst brauche ich jedoch noch personenbezogene Daten. Keine Sorge die Daten werden anonymisiert.

(Drücke die Leertaste um fortzufahren.)

(Introduction screen 2)

Im Folgenden wirst Du einen Namen sehen sowie einen 7-stelligen Slider. Du bewertest den Namen von ganz links "sehr männlich" bis ganz rechts "sehr weiblich". Die Mitte der Skala zeigt "neutral" an.

Zum Beispiel würde ich "Uwe" als "sehr männlich", "Gudrun" als "sehr weiblich" und "Alex" als "neutral" bewerten.

Wichtig ist, NICHT die assoziierten Qualitäten des Namens zu bewerten (Gudrun klingt hart/stark also "männlich"), sondern ob Du eher an einen Mann (der Uwe), eine Frau (die Gudrun) oder vielleicht beide (der/die Alex) denkst.

Three practice items, which did not appear in the rest of the experiment, were presented such that participants could accommodate themselves with the procedure. A final screen stated that the experiment would start now. The 143 names and the rating scale were shown one by one in the centre of the screen (for visualization see Figure 3.1) with an *interstimulus interval* of 500 ms. The rating scale was labelled "sehr männlich" (*very male*) (leftmost) to "neutral" (*neutral*) (centre), to "sehr weiblich" (*very female*) (rightmost). The remaining four steps were unlabeled, so the participants could interpret the space themselves, ensuring that words like "eher" (*rather*), "tendenziell" (*tentatively*), or "hauptsächlich" (*mainly*) do not cause biases (Shinar, 1975: 101). During the reading and rating period, the participants were not time-restricted.

Finally, the participant's gender ("männlich" (*male*), "divers" (*non-binary*)), "weiblich" (*female*)) and gender identity (7-point slider from "männlich" (leftmost) to "weiblich" (rightmost)) were inquired after the experiment such that the gender inquiry does not interfere with the participant's neutral mindset during the task.

4.1.4 Selection Criteria

Previous rating studies (Carreiras et al., 1996; Irmén, 2007; Irmén & Kurovskaja, 2010; Kennison & Trofe, 2003; Shinar, 1975) have taken the *mean* rating as a measure of gender association. Since I investigate ambiguous names, *mean* ratings alone have the downside that if half of the sample population rate a name as “sehr männlich” (1 on the rating scale), and the other half as “sehr weiblich” (7 on the rating scale) the unanimously unambiguous name would be classified as ambiguous (4 on the rating scale). Consequently, I also used the *median* rating as an additional passing test for the name–gender association and calculated the *mean’s standard deviation (sd)* for another comparison.

Definition of an objective name–gender association

- A name with a mean rating between 1 and 1.3 and a median rating of 1 or 2 is defined as male
- A name with a mean rating between 3 and 5 and a median rating between 3 and 5 is defined as ambiguous
- A name with a mean rating between 6.7 and 7 and a median rating of 6 or 7 is defined as female

4.2 Results and Discussion

The final list of rated names consisted of:

- 26 male names
- 24 ambiguous names
- 27 female names

Figure 4.2 shows for each of the 35 participants the mean gender rating of each name on the x-axis and the participant gender rating on the y-axis. Each dot represents one name in the two-dimensional rating space while the colour (red: male; green: ambiguous; blue: female) represents the potential gender – the gender postulated by the Google query. Figure 4.2 indicates that inter-participant gender ratings differ between participants, and the participants could generally be grouped into three categories. VP03 and VP33 use the whole rating scale range, especially for the potentially ambiguous names. VP07, on the other hand, strictly classifies names on the extremes of the scale. Lastly, VP11 and VP14 use the extremes but also the centre of the scale – the labelled sections on the rating scale.

In response to a familiarity judgement task Valentine et al. (1991: 164) write, “[i]t is impossible to discuss ‘error rate’ because it is possible that a subject responded ‘no’ to a name rated

as familiar because the name was genuinely unfamiliar.” I suggest the same for the name–gender association rating. The name–gender association rating of a participant might reflect that individual’s belief system. No participant was rejected since no participant seemed to work directly against the task.

Figure 4.2: A dot-plot depicting mean gender rating on the x-axis by subjective gender rating on the y-axis for each participant ($n = 35$). The colour represents the potential gender which was determined by the Google queries

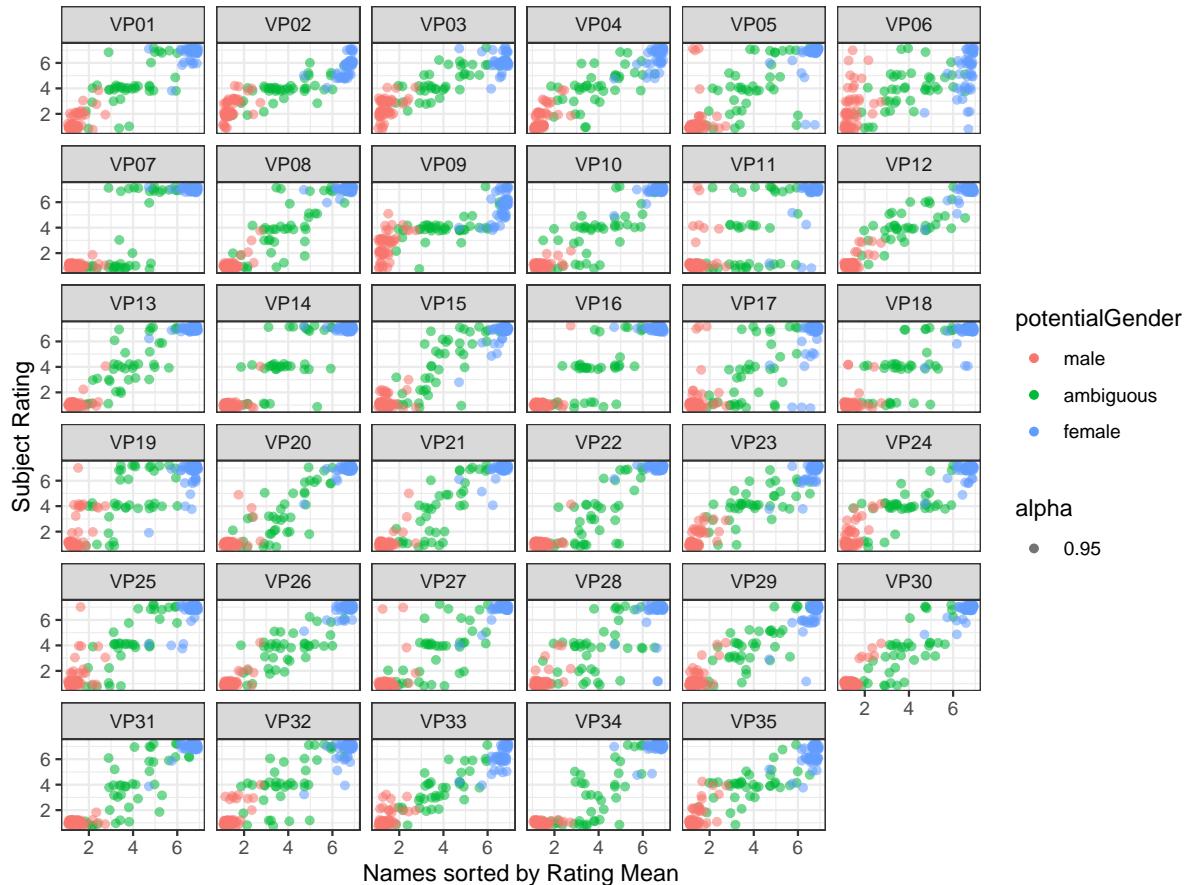
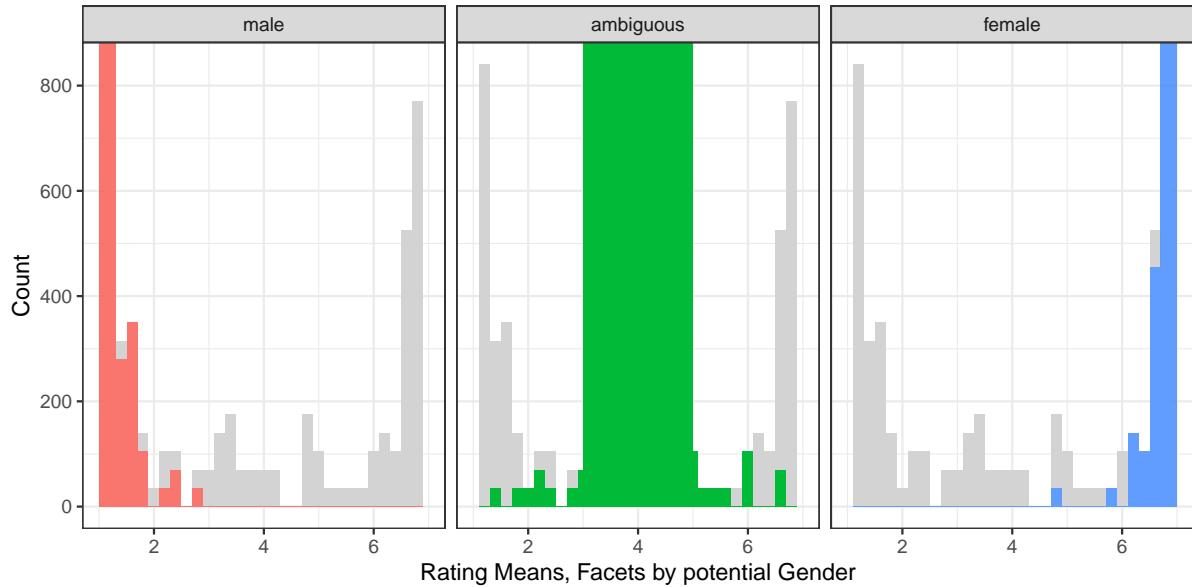


Figure 4.3 shows the spread of the mean gender ratings broken down by potential gender – the Google classification. The rating means are displayed on the x-axis and counts are stacked on the y-axis³. The colours match the potential gender described above. The highlighted areas (rectangles) encircle all names that are for the selection criteria “mean rating” male, female or ambiguous. Potentially male names (red) and potentially female names (blue) are rated so that their spread is very narrow. Potentially ambiguous names, on the contrary, have mean ratings spread across the whole range of the rating spectrum. Four names postulated by the websites as potentially ambiguous were rated very male or female, and one potentially female name fell into the objectively ambiguous rectangle. This show that a Google search query, especially for ambiguous names, does not align with an objective rating study.

³The bin width was set to 0.2.

Figure 4.3: Mean gender rating split by potentially male (red), ambiguous (green), and female (blue) names. The rectangles describe a selection criterium and the grey shade shows all data in all facets



Lastly, Figure 4.4 compares Kennison and Trofe, 2003's role name ratings with the rating of proper names from my study. The mean gender rating of said items is plotted against their standard deviation (sd). Kennison and Trofe (2003) write, “the most agreement as reflected in low standard deviations exists for items at each end of the scale and also at the middle of the scale. This pattern indicates the following: (i) there are items that are viewed as referring mostly to females, with high agreement across participants; (ii) there are items that are viewed as referring mostly to males, with high agreement across participants; and (iii) there are items that are viewed gender neutral, with high agreement across participants.” I suggest there is a fourth observation to be made. Some role names, indicated by their high sd , are interpreted by some as more male or more female, forming a group of “undecidedly ambiguous” role names. For proper names, the (third) group “agreeably gender ambiguous” does not exist but all ambiguous names rather fall into the “undecidedly ambiguous” group. There is no ambiguous name with a sd lower than 1. Even though most individuals (see Figure 4.2) classify many names as “neutral”, classifying names as neutral or ambiguous for this and future sample populations is difficult to justify.

This highlights, on the one hand, role names and proper names are in terms of their “agreeably gender ambiguous” class different. Hence, comparisons between studies with role names and studies with proper names as stimuli are not straightforward. On the other hand, the name–gender association for neutral/ambiguous role names and proper names has high inter-participant variation. This suggests, a post hoc rating task for each participant should be carried out so that the Main study does not rely on the potentially gender-misaligned name–gender associations superimposed from the Norming study on the participants.

Figure 4.4: A comparison of the distribution of mean gender ratings for role names or proper names by their standard deviation

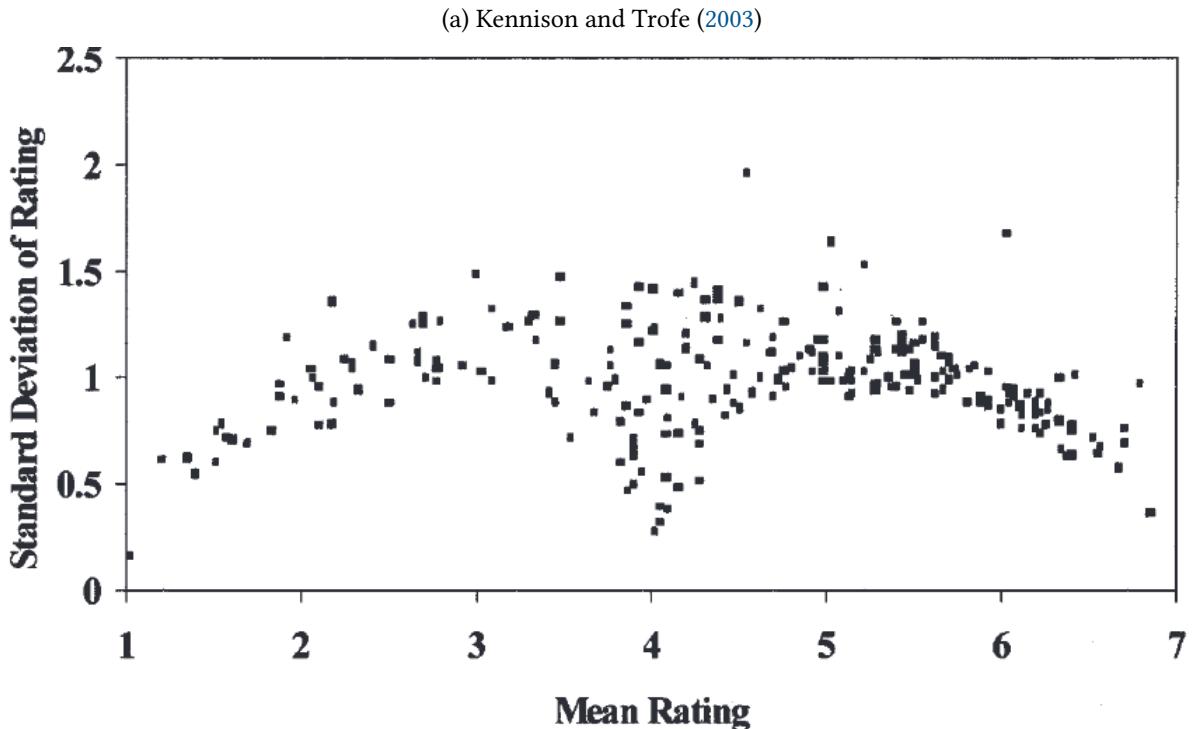
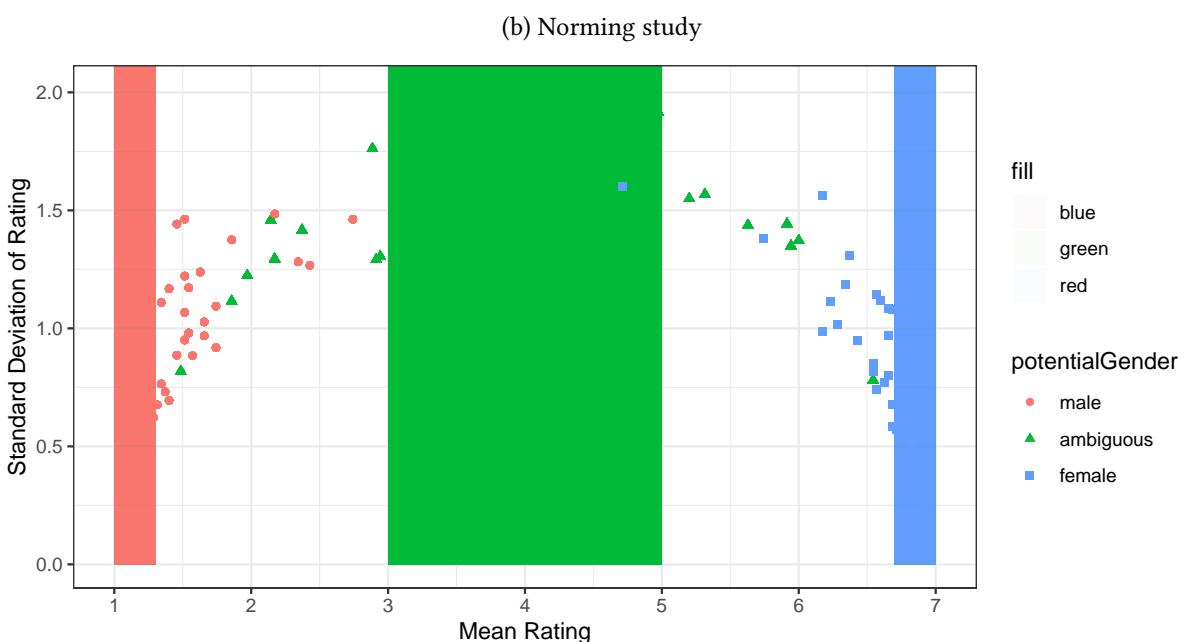


Fig. 1. Scatterplot of mean rating for each item by standard deviation for each item from Study 1.



Chapter 5

Main and Post Hoc Study

This thesis tries to show that there is more to anaphora resolution with proper names than match and mismatch. The Norming study has established that there is a need for a subjective name–gender association, so this chapter consists of two interlocking studies. First, I will present how I designed and conducted a self-paced reading study in the style of Carreiras et al. (1996) and Kennison and Trofe (2003) but used first names that are objectively male, female or ambiguous. Second, I will describe the Post Hoc name–gender rating study, which allowed for predictor variables that capture gender information similar to the Ackerman (2019) framework. Lastly, I will present relevant results that might spark a discussion about the present state of stimuli classification.

5.1 Main Study

5.1.1 Participants

60 (i) German native speakers, (ii) between the age of 18 and 35, (iii) naïve to the purpose of the present experiments and the Norming study, took part in two online studies spaced two weeks apart. Due to a bug in the first experiment, four participants needed to be removed because 20% of their data was not recorded. Five participants, obeying selection restrictions (i) – (iii) stated above, were added to the subject pool, such that the final sample population consisted of 61 participants (32 females, 22 males, 2 non-binary individuals, and 5 who did not specify their gender; mean age: 25.30 years (*range*: 19 – 35 years, *sd*: 3.81 years)). All 66 participants were compensated 7 Euros.

5.1.2 Materials and Design

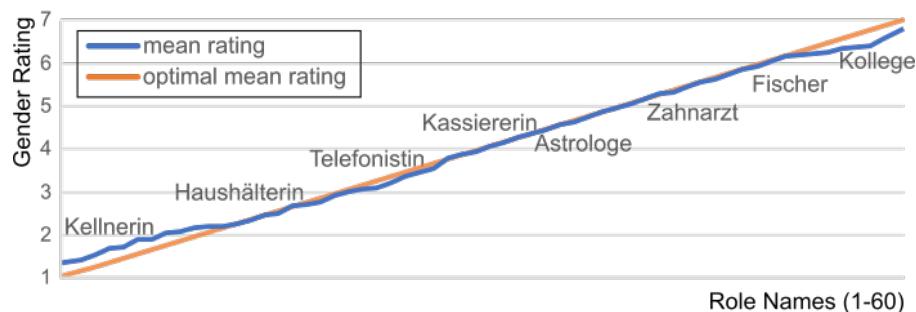
Items The Norming study provided 60 target items (20 male names, 20 female names, 20 ambiguous names) as well as 72 names serving as alternative comprehension question answers or warm-up items. From Kennison and Trofe (2003)'s 405 role names, I chose 60 filler items

in such a way that they have an optimal smooth transition from “strongly male” to “strongly female” (see Figure 5.1) while obeying selection rules stated below¹. I included the smooth transition to ensure that there are “ambiguous-like” role names (e.g. *Künstlerin* (*artist*) or *Psychater* (*psychiatrist*)), even if the stereotypical gender of *Künstler* might be overwritten by the grammatical gender of -*in* (Carreiras et al., 1996) or if the German generic masculine role names (e.g. *Psychater*) are perceived like the explicit masculine role names demonstrated by Schmitz (2022). I am also aware of the fact that the stereotypical gender rating captured in the USA in 2003² differs from the stereotypical gender rating in Germany in 2022³, but these 60 role names served only as fillers.

Selection criteria for role names (filler items):

- No Role Name with definitional gender (e.g. *Vater* (*father*) or *Krankenschwester* (*nurse*))
- No Role Name without a male/female counterpart (e.g. *Kumpel* (*dude*))
- Only Role Names that only take the suffix -*in* to form the female form (e.g. no *Bediensteter*/*Bedienstete* (*official/official*))

Figure 5.1: The distribution of filler items and their mean gender rating taken from Kennison and Trofe (2003) vs the optimal distribution of mean gender rating



Sentences The sentence pairs were constructed as illustrated in Example 5.1 and presented word-by-word, with the exception of region 03 (reg 03), illustrated in Figure 5.2. Following Kennison and Trofe (2003) and Nieuwland et al. (2007), I used a mixture of different themes determined by the location (PP) of the first sentence to increase content variability, but at the same time, following a template similar to Kennison and Trofe (2003) (see row two of Example 5.1). The subject (item and pronoun) was locked to the sentence-initial position. The complex PP at presentation region 03 was presented as a whole for fear of ungrammaticality effects which might have spillover effects on the target word. Lastly, no relevant region was placed

¹I calculated the “optimal mean rating” – a linear transition – from 1 to 7 for 60 items and selected the role names that had the closest mean gender rating to that variable.

²This is the place and time of Kennison and Trofe, 2003’s data collection.

³Time of data collection.

at the end of a sentence to avoid inference from sentence wrap-up effects (Just & Carpenter, 1980: 331). The regions of interest were the item and its spillover region (reg 01 and reg 02) as well as the pronoun and its three spillover regions (reg 04 to reg 07).

Table 5.1: An example sentence, its chunks, the regions of interest, and the region (reg) number. “Pronoun” is shortened to “Pro” and $x+n$ indicates how many regions (n) the present region is presented after x

Alex	guckt	aus dem Fenster.	Sie	hat	einen guten Freund	gesehen.
Item	V	PP	Pronoun	AUX	DET	ADJ N V
Item	Item+1		Pro	Pro+1	Pro+2	Pro+3
reg 01	reg 02	reg 03	reg 04	reg 05	reg 06	reg 07 reg 08 reg 09

The first sentence was created from a semi-random combination of verbs and PPs drawn from the publicly available website Deutschlernerblog (2019). I wrote the second sentence in such a way that it (i) obeyed the template and (ii) fit the context of the first sentence. All 131 carrier sentence pairs (for the 60 target items, 60 filler items, and 11 warm-up items) were different to ensure that no sentence pair could prime the interpretation of a subsequent name (van Dijk & Kintsch, 1983: 11) (i. e. “I know that there was Phillip, a man, at the bus stop 3 sentence pairs ago, so Alex is more likely to be a man, too”). A list of the carrier sentence pairs can be found in Appendix A. Subsequently, items and pronouns were combined with the carrier sentence pair in such a way that item-gender (male, female, and ambiguous) and pronoun-gender (male, female) were fully crossed, resulting in a 3×2 within-subject design. For each combination, 10 sentences were presented (See Table 5.2 for an overview of the six lists). Each list was randomized so that *PsychoPy3* was set to “sequential order”.

Table 5.2: Item-Pronoun combination in carrier sentences throughout six lists; **M** = Stereotypical Male Item, **F** = Stereotypical Female Item, **A** = Ambiguous Item, m = masculine Pronoun (dark colour), f = feminine Pronoun (light colour)

Sentences	List					
	1	2	3	4	5	6
1 – 10	M_m	M_f	F_m	F_f	A_m	A_f
11 – 20	M_f	M_m	F_f	F_m	A_f	A_m
21 – 30	A_m	A_f	M_m	M_f	F_m	F_f
31 – 40	A_f	A_m	M_f	M_m	F_f	F_m
41 – 50	F_m	F_f	A_m	A_f	M_m	M_f
51 – 60	F_f	F_m	A_f	A_m	M_f	M_m

Note: The table only displays target items.

Of each sentence pair, six versions were created, which were spread over six lists. In version 1, the subject of the first sentence was a stereotypically male name (e. g. *Phillip*)

and the subject of the second sentence was the gender-matching pronoun *er*. In version 2 of the sentence pair, the pronoun was the gender-mismatching *sie*. In versions 3 and 4, the sentence pair was constructed with a stereotypically female name (e.g. *Anna*) and the then gender-mismatching pronoun *er* and gender-matching *sie*. The sentence pair in versions 5 and 6 contained ambiguous names (e.g. *Alex*) for which it is unclear whether *er* and *sie* are matching or mismatching because there should be valid and invalid referents (*Alexander* or *Alexandra*) in the parsers mental lexicon. Since the association of gender and names differs from participant to participant, the Post Hoc study was undertaken.

Comprehension Questions I compiled a forced-choice comprehension question for each sentence pair with one true and one false answer (Examples are stated in Table 5.3). Their aim was to distract the participants from the experiment's goal, increase the experiment's demand, and provide a measure of the participant's attentiveness. The comprehension questions targeted all regions except for regions 04, 05, and 09. I wanted the most natural reading procedure for regions 04 (pronoun) and 05 (first spillover zone) so they were not targeted by the questions. Region 09 was often liked to region 05 due to the use of predicative verb constructions (e.g. *hat ... gesehen*). Targeting regions 06 to 07 ensured that participants also concentrated on the content of the second sentence. As in Kennison and Trofe (2003: 362), the comprehension questions specifically did not refer to “the sex of the individuals described in the sentences”.

Table 5.3: An example of comprehension questions for each targeted region (reg)

Example Question	region	Sentence / Answer	Alternative Answer
Wer guckt aus dem Fenster?	reg 01	Alex	Anna
Was tat Alex?	reg 02	gucken	schauen
Wohin guckt Alex?	reg 03	aus dem Fenster	in das Büro
—	reg 04	Er	
—	reg 05	hat	
Wen hat Alex gesehen?	reg 06-08	einen guten Freund	einen guten Kommilitonen
—	reg 09	gesehen	

Note: Other alternative answers for “Wen hat Alex gesehen?” were “den guten Freund” or “einen geliebten Freund”.

5.1.3 Procedure

The participants were sent an email with their participant number, reasons why the personal information is important to ensure that they truthfully provide information and a link to their list in *Pavlova*. While they filled in their information (participant number, email address, first name, last name, age, handedness, and up to five L2s), their computer downloaded the experiment data such that internet bandwidth fluctuations did not affect the procedure of

the experiment. After a welcome screen, two introduction screens explained the experiment procedure.

(Introduction screen 1)

“Im folgenden wirst du je zwei Sätze Stück für Stück lesen.

Das heißtt Du wirst immer ein Wort oder ein Satzstück sehen und Du entscheidest, wann du das nächste Stück sehen möchtest. Du kommst immer mit der Leertaste zum nächsten Wort/Satzstück. Es gibt keine Möglichkeit um zurückzukehren.

Drücke nun die Leertaste um weiterzukommen.”

(Introduction screen 2)

“Manchmal wirst du eine Frage mit zwei Antwortmöglichkeiten über die Sätze beantworten. Lies die Sätze also möglichst genau aber auch so schnell wie möglich. Es ist immer nur eine Antwortmöglichkeit korrekt.

Die obere Pfeiltaste wählt die obere Antwort aus und die untere Pfeiltaste wählt die untere Antwortmöglichkeit aus.

Ca alle 3 bis 5 Minuten hast Du eine Pause.

Jetzt folgen fünf Beispiele an denen Du die Steuerung testen kannst.

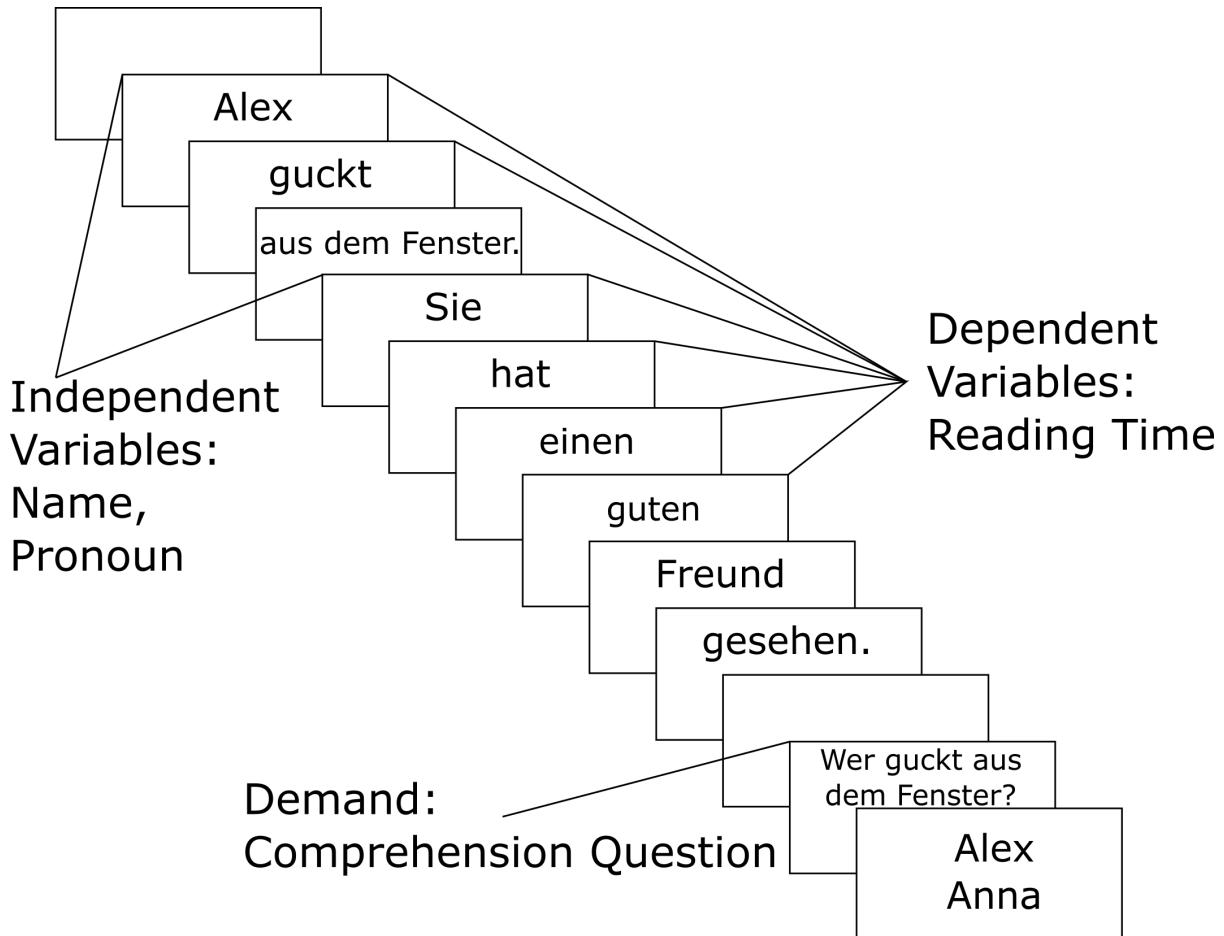
(Drücke die Leertaste um weiter zu kommen.)”

After the introduction screens, they were presented with five practice trials followed by a screen that reminded them to remain undisturbed, have a beverage in arms reach for the breaks and that the experiment would start now. As illustrated in Figure 5.2, the sentence pairs were presented word-by-word⁴ in the centre of the screen (*PsychoPy3* settings: Font: Open Sans, Letter height: 0.1). The participant pressed the space bar to move to the next presentation region. There was no timeout. Forced choice comprehension questions (Letter height: 0.08) with the answer on a subsequent screen⁵ were shown at randomly predefined 25% of the sentence pairs. The *interstimulus interval* was set to 500 ms, and roughly every 3 to 5 minutes, there was a break without a time limit. At the start of the experiment, there were six warm-up sentence pairs, and after each break, the experiment continued with one dummy sentence. None of the warm-up/dummy items was shown in the experiment again. At the end of the experiment, the participants were reminded that there would be a second experiment in two weeks. The first experiment lasted approximately 25 minutes.

⁴Except for region 03 for reasons stated above.

⁵The participants used the arrow keys *up* or *down* to choose either option. Using *up/down* was preferred over *left/right* due to screen space limitations.

Figure 5.2: Procedure of the self-paced-reading experiment. The comprehension questions were shown 25% of the time



5.2 Post Hoc Study

To understand every participant's subjective name–gender association, I conducted a post hoc rating study.

The participants ($n = 66$) were identical to the Main study, and the material (143 names) and the design (7-point rating scale, $ISI = 500\text{ms}$) were identical to that in the Norming study (Chapter 4 for reference). The only difference in the procedure was that participants were not asked for personal information, and the introduction texts were slightly different. “Introduction screen 1” was omitted, and “Introduction screen 2” was split into two screens because more text was added (see below). All 66 participants received the same link, and the experiment lasted approximately 10 minutes.

(additional text)

Klicke mit der Maus auf einen der sieben Punkte an dem du denkst, wo sich der Name deiner Meinung nach befindet.

Nun Folgen drei Beispiele, an denen Du die Steuerung testen kannst.

I did not conduct an analysis for the Post Hoc study data similar to that of the Norming study because the purpose of the Post Hoc data lies in the combination with the Main study's data.

5.3 Statistical Analysis

5.3.1 Calculating the Main Predictors of Theoretical Interest

The data from the Post Hoc study was pre-processed and merged with the data from the Main study with *Python* (version: 3.11.2; van Rossum and Drake (2009)) using the *pandas* (version 2.0.0; McKinney (2010) and The *pandas* development team (2020)) library while data cleaning and the statistical analysis was done with *R*.

One of the main questions in the statistical analysis was: How can I measure gender mismatch in the Ambiguous condition? The gender of *er* or *sie* either always⁶ or never⁷ matches an ambiguous name. To answer this question, I used the difference between the item rating and the pronoun rating value. The gender rating from the Post Hoc study provided for every name a participant subjective gender rating (*Item.Rating*) from 1 ("sehr männlich") to 7 ("sehr weiblich"). I propose that the grammatical gender of the pronouns *er* and *sie* are binary since *es* is generally not used in reference to (non-diminutive) humans and, as such, should be equated to the extremes of the rating scale (*Pronoun.Rating*). So *er* has the value 1 and *sie* has the value 7. The absolute value of the difference between the two variables (*Item.Rating* and *Pronoun.Rating*) is the *participant_itemPro_mm* (i. e. the participant-specific mismatch between item and pronoun rating).

$$\text{participant_itemPro_mm} = |\text{Item.Rating} - \text{Pro.Rating}|$$

To form Match, Ambiguous, and Mismatch conditions (*participant_mm_grouping*) from the gender mismatch gradient *participant_itemPro_mm*, I used the gender boundaries stated in the formula below. An example in Table 5.4 illustrates the importance of the main predictor of interest *participant_mm_grouping*.

$$\begin{aligned} 0 \geq \text{participant_itemPro_mm} \geq 1 &\rightarrow \text{participant_mm_grouping} == \text{"Match"} \\ 2 \geq \text{participant_itemPro_mm} \geq 4 &\rightarrow \text{participant_mm_grouping} == \text{"Ambiguous"} \\ 5 \geq \text{participant_itemPro_mm} \geq 6 &\rightarrow \text{participant_mm_grouping} == \text{"Mismatch"} \end{aligned}$$

Two participants have rated *Anna* a 7 (female), but the pronoun presentation differed. This results in the case that for participant 01 *Anna...sie* is in the Match condition, and for participant 02 *Anna...er* is in the Mismatch condition. Further, participant 01 rated *Alex* a 4 (neutral/ambiguous) while participant 02 rated *Alex* a 1 (male) such that for participant

⁶*Alex* is a name with two genders.

⁷*Alex* is underspecified.

01 *Alex...sie* is in the Ambiguous condition, and for participant 02, despite having the same anaphora and antecedent as participant 01, *Alex* is in the Mismatch condition. This shows that the variable `participant_mm_grouping` is participant specific and allows for conditions that are a-typical for the present research.

Table 5.4: Potential results from Post Hoc study and their incorporation with the Main study results evaluation

Participant	Item	Item.Rating	Pro	Pro.Rating	Difference	Grouping
01	Anna	7	sie	7	$ 7 - 7 = 0$	Match
01	Phillip	2	sie	7	$ 2 - 7 = 5$	Mismatch
01	Alex	4	sie	7	$ 4 - 7 = 3$	Ambiguous
02	Anna	7	er	1	$ 7 - 1 = 6$	Mismatch
02	Phillip	1	er	1	$ 1 - 1 = 0$	Match
02	Alex	1	sie	7	$ 1 - 7 = 6$	Mismatch

5.3.2 Data Cleaning

The initial data frame consisted of 8,768 rows and 50 columns. On average 93% of the comprehension questions were answered correctly, and no participant fell below the exclusion threshold of 75%⁸, indicating that participants read the sentence pairs carefully and understood the task. The reading times across the presentation regions ranged from 1.20 ms to 57,837.90 ms, so I manually inspected the summed reading time for every participant and carrier sentence. None were rejected because there were large variations throughout.

The reading time (rt) distribution did not pass the Shapiro-Wilk-Test (Shapiro & Wilk, 1965) for normality, so the data frame was cut due to the striking number of outliers. Three methods were chosen, and later AIC comparisons should decide which data frame (df) I would continue using. (1) `df_t` was filtered for target items but otherwise uncut. (2) In `df_cat`, the rts, per participant, of all nine regions were summed up, and each sentence pair which had a summed-rt outside the *interquartile range* (*IQR*) was removed (4.40% of data lost). (3) In `df_5k`, the whole sentence pair (nine regions) was removed if any rt in the nine regions was shorter than 200 ms (Baayen, 2011: 265) or longer than 5000 ms. Additionally, the rts in `df_5k` were trimmed with $\pm 2.5sd$ per region per participant (23.09% data lost). (4) `df_2k` went through the same procedure as `df_5k`, but the upper bound was reduced from 5000 ms to 2000 ms (27.30% data lost) in order to stay within a reasonable conservative range. The high loss of data stems from the fact that the whole sentence pair was removed, given one region violated said bounds.

All methods failed the Shapiro-Wilk-Test, so I consulted the `bestNormalize` (Peterson,

⁸Participant 38 did not pass the text but was one of the four participants that were rejected due to a bug in the experiment.

Table 5.5: List of potential predictor variables, their type, and values. Crossed-out variables were not included in the best model

Predictor Variables	Variable-Type	Value
participant_mm_grouping	factor	“Match”, “Mismatch”, “Ambiguous”
participant_itemPro_mm_num	numeric	0, 1, 2, 3, 4, 5, 6
list	factor	“1”, “2”, “3”, “4”, “5”, “6”
trial_index_z	numeric	range: -1.837250, 1.681417
pro (pronoun)	factor	“er”, “sie”
item_freq_z	numeric	range: -1.3840592, 1.90560987
participant	factor	“1” – “85”
participant_gender	factor	“m”, “f”, “nb”, “na”
participant_age_z	numeric	range: -1.6998338, 2.4885567
item_id (proper name)	factor	“1” – “72”
block	factor	“1”, “2”, “3”, “4”, “5”, “6”
sent_id (carrier sentence)	factor	“1” – “60”
item_gender_norming	factor	“female”, “male”, “ambiguous”
handedness	factor	“lefthanded”, “righthanded”
L2	factor	“fra”, “jpn”, “ita”, ...

2021) package and normalized the reading times accordingly with the *Ordered Quantile (ORQ)* normalization transformation (Peterson & Cavanaugh, 2020). The data was still not normally distributed ($W = 0.89847$, $p < 2.2\text{e-}16$), but other methods of data transformation exceeded the scope of this thesis.

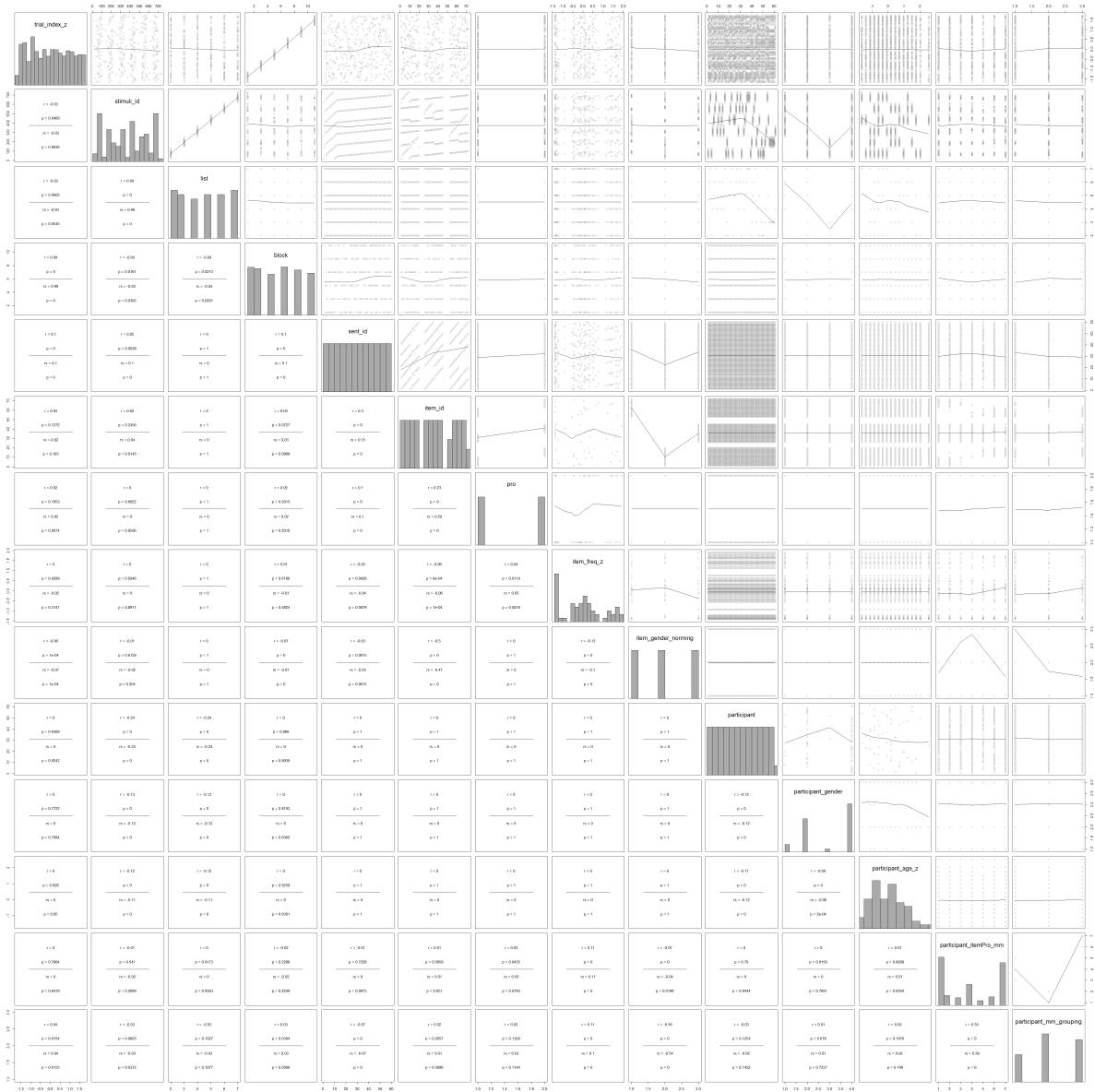
5.3.3 Fitting the Best Model

Avoiding Colinearity

Before fitting models, I ensured that no colinearity effects would affect the model predictions. I constructed a correlation matrix (see Figure 5.3) on df_t and removed one of two variables that had a Pearson correlation coefficient r , for numeric variables, or a Spearman correlation coefficient rs , for discrete variables, greater than 0.3. A full list of predictor variables is stated in Table 5.5. Crossed-out variables are those that did not end up in the best model.

block ($rs = 0.99$), sent_id ($rs = 0.31$), and item_gender_norming ($rs = 0.43$) were rejected, because they correlated with the more promising predictors trial_index_z, item_id, and participant_mm_grouping. trial_index_z is a more fine-grained representation of the same data than block, item_id is more important than sent_id because it was planned as a random effect, and participant_mm_grouping was the main predictor of interest, which is more important than item_gender_norming. handedness was neglected in pre-processing since only male participants were left-handed, leading to an implicit gender bias in the predictor variable. L2 was not included because Dummy Coding would have inflated the model with 26 different L2s. Also, the main predictors of theoretical interest

Figure 5.3: Correlation matrix of all potential predictor variables



(`participant_mm_grouping` and `participant_itemPro_mm_num`) correlate, but they are never used in the same model. The explanation for the elimination of the random effect `item_id` follows below.

From Full Model to Best Model

I fitted a linear mixed effects regression model using the `lme4` (Version 1.1-32; Bates et al. (2015)) package for each of the six regions of interest (region 01, region 02, region 04, region 05, region 06, region 07) times the four types of data frame cuts (`df_t`, `df_cat`, `df_5k`, `df_-2k`). To deduce the “best model”, I used top-down stepwise regression with AIC comparisons⁹.

⁹For the Akaike Information Criterion (AIC), a lower value indicates a better model fit. The delta symbol (Δ) can be read as ‘change in’.

I decided not to remove predictors that were not significant but rather stop removing them once the AIC value was at its lowest.¹⁰ The AIC comparisons identified (i) which method of data frame subsetting is the best, (ii) which fixed effects and (iii) which random effects were worth keeping. AIC comparison batteries (i) and (ii) were tested on the most relevant presentation region 04 (pronoun), while (iii) was tested on all six regions because especially `item_id` should be sensitive at region 01 (item) and 02 (item spillover).

I followed Baayen (2011: 279)'s $\pm 2.5sd$ residual trim instructions as an attempt to normalize the data and fit for each of the four data frames (`df_t`, `df_cat`, `df_5k`, `df_2k`) an `lmer` model with a residual trim and one without (see Example 5.3.3). Subsequently, I computed AIC values for the six models to check the model fit (see Table 5.6). The comparison indicates, on the one hand, the more cut the data frame is, the better the model fit, and on the other hand, the residual trimmed models fit their data better than those without the residual trim. `mdl_2k_P4_trimmed` has the overall best model fit with an AIC value of 31339.14.

After I selected the best data frame (`df_2k`), I iteratively removed variables to move from a full model to the best model. The comparisons showed that the model fit does not improve when any fixed effects are removed (see Table 5.7), but the fit improves from AIC = 31339.14 to AIC = 31337.46 when the random effect `(1 | item_id)` is removed. I calculated the impact of removing `(1 | item_id)` for all regions (see Table 5.8), and the random effect improves the model on average by .05% and .21% at best, so a simpler model is preferred. Another round of AIC comparisons was calculated, but the model fit did not improve if any other predictor is removed such that the residual trimmed model displayed below (e.g. `mdl_best_R4_trimmed`) is the “best model”.

Table 5.6: AIC values for lmers with different data frames at the most important region

Rank	model, data frame (df), trimmed (yes/no)	AIC at R4	Δ Rank 1
1	full Model, df = 2k, residuals trimmed	32228.55	0.00
2	full Model, df = 2k	33232.60	1004.05
3	full Model, df = 5k, residuals trimmed	33945.55	1717.00
4	full Model, df = 5k	35051.39	2822.84
5	full Model, df = cat, residuals trimmed	45167.62	12939.07
6	full Model, df = cat	46218.17	13989.62
7	full Model, df = full, residuals trimmed	48867.85	16639.30
8	full Model, df = full	49747.61	17519.06

¹⁰Baayen, 2011: 259 and Winter (2019: 277) agree that there is no “one strategy” to deduce the best model or and state that the strategy should be discouraged from if there is a more sensible approach. Still, it is the norm in linguistic literature, and for a novice researcher, it might be the most sensible entrance to good model selection.

Table 5.7: AIC comparisons between full model, and models with fewer predictor variables

Rank	Model	AIC R4	Δ full Model
1	– (1 Item)	32227.50	-1.05
2	full Model	32228.55	0.00
3	– Participant Age	32234.34	5.79
4	– Pronoun	32235.86	7.31
5	– Participant gender	32258.32	29.77
6	– List	32269.20	40.65
7	– Trial index	32373.91	145.36
8	– Item frequency	33278.83	1050.28

Table 5.8: AIC comparisons between full model and model without item as random effect

	AIC R1	AIC R2	AIC R4	AIC R5	AIC R6	AIC R7
Δ full Model	72.25	23.61	10.91	9.84	-2.00	-13.73
– (1 Item)	34016.72	31337.56	32232.34	29671.97	29871.55	31287.61
full Model	33944.47	31313.56	32221.43	29662.13	29873.55	31301.34

“Best Model”:

```
mdl_best_R4 <- lmer(rt_pos04_ordNorm ~ participant_mm_grouping +
(1 | participant) + trial_index + list + pro + item_freq_z +
participant_gender + participant_age_z, df_2k)

mdl_best_R4_trimmed = lmer(rt_pos04_ordNorm ~ participant_mm_grouping +
(1 | participant) + trial_index + list + pro + item_freq_z +
participant_gender + participant_age_z, df_2k,
subset = abs(scale(resid(mdl_best_R4))) < 2.5)
```

Model Criticism

When working with linear mixed-effects regressions, three assumptions are tested so that we can trust that the model is reliable enough to make predictions about the population. We assume the residuals are normally distributed, the fitted values and the residuals have a linear relationship, and their variance is unbiased.

Figure 5.4 shows the distribution of the model’s residuals in blue and a normal distribution in green. The residuals are negatively skewed, have a heavy “fat tail” on the right and are visually not normally distributed because the blue density distribution does not follow

the green normal distribution. The function `check_normality` also indicates that the Non-normality of residuals is highly significant ($p < .001$). Winter (2019: 110) checks normality via a *Q-Q plot*. Figure 5.5 reveals that the Quantiles from the sample data do not conform with the Quantiles of a normal distribution. The left tail deviates from the normal distribution (the straight line), starting at -2 sd . The sample quantiles diverge heavily from the quantiles of a normal distribution starting at 1 sd , indicating the heavy “fat tail”.

Figure 5.5(b) reveals that the constant variance assumption or homoscedasticity, the same variance in all conditions (Brehm & Alday, 2022: 3), is not met (Winter, 2019: 109). The data is not clustered randomly but instead funnels strongly. Figure 5.6 also suggests non-constant variance due to the funnelling described before. To underpin these visual claims, the function `check_heteroscedasticity` clearly states a significant detection of heteroscedasticity ($p < .001$) (i. e. the opposite of homoscedasticity).

Figure 5.6 also indicates that the linearity assumption is not met since the data clearly does not show a linear and horizontal relationship between the fitted values and the square root of absolute residuals. I can not assume a linear relationship between the predictor and predicted variables.

In sum, the “best model” does not fulfil the three critical assumptions that should be fulfilled to abstract the significant insights from the model to the population. The plots suggest that the model does not capture important non-linear relationships in the data (skewness and non-linearity) and that the model does not account for extreme values (tailedness and heteroscedasticity) even though the data has been cut, normalized and trimmed (see Section 5.3.2). An objection to the validity of the model assumption is posed by Schielzeth et al. (2020) and will be discussed in Section 6.3. I repeated the process of model criticism for the second main predictor of interest, `participant_itemPro_mm`, which interprets gender mismatch as a gradient instead of three discrete categories. The results are identical – all three assumptions are not met. The figures can be found in Appendix Figures B.1a to B.1c.

Figure 5.4: Normal probability plot for the final model at region 04

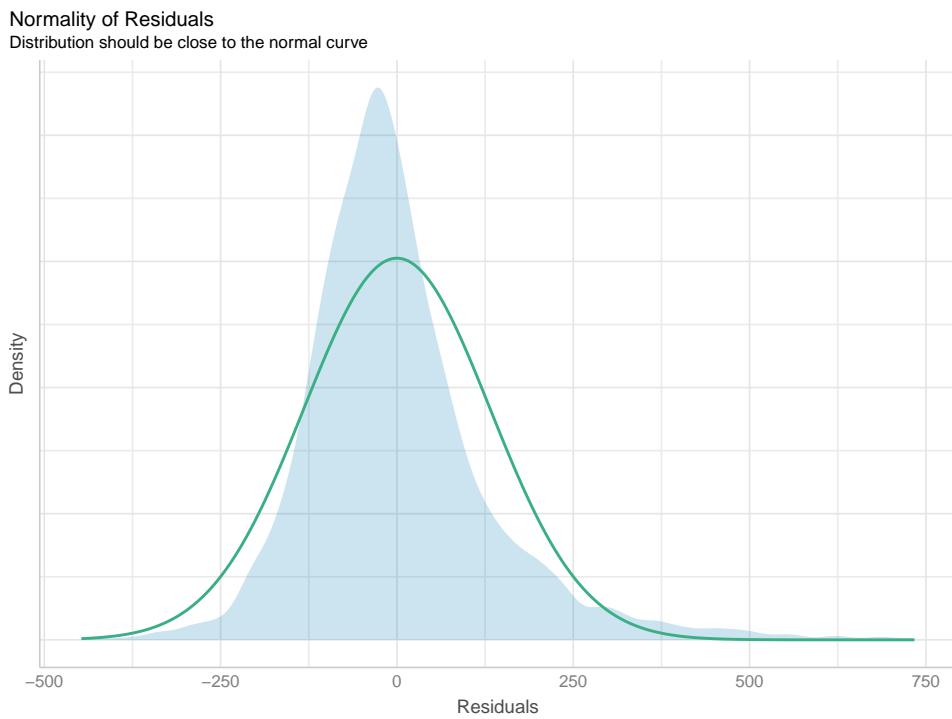


Figure 5.5: Q-Q plot and residual plot for the final model at region 04

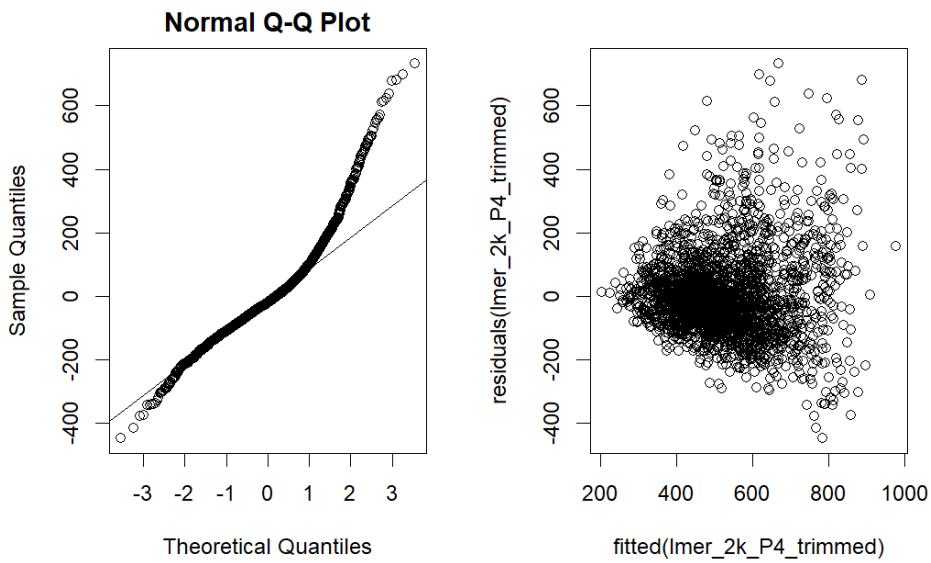
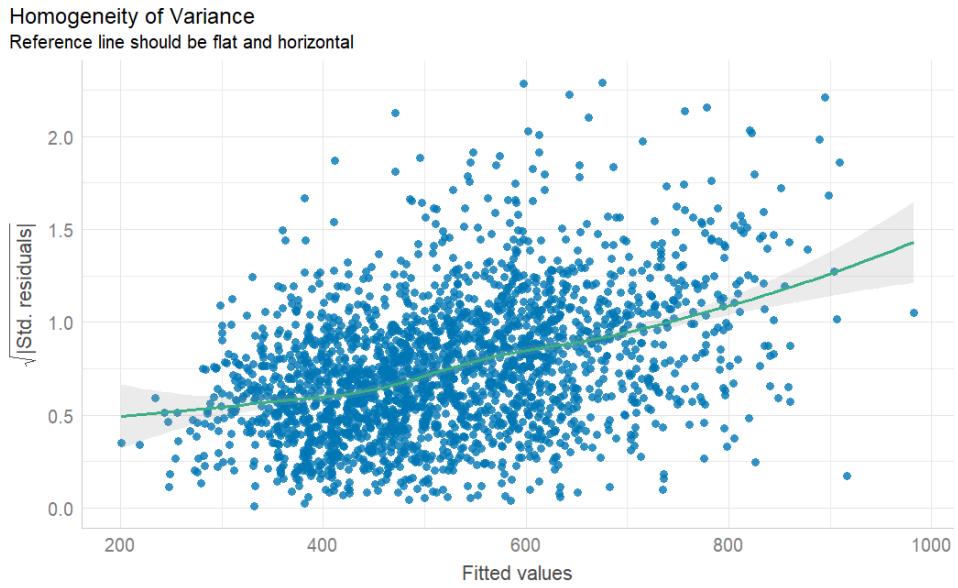


Figure 5.6: Linearity and Homoscedasticity plot for the final model at region 04



5.4 Results

5.4.1 Categorical Analysis

Following the “best model”, all analyses used the fixed effects: trial index, list, pronoun, item frequency, and participant’s gender and age. In the models for regions 01 and 02, the fixed effect Mismatch grouping had the values “NonAmbiguous”¹¹ and “Ambiguous” since either Match or Mismatch cannot be observed before a gender agreeing or disagreeing anaphora is presented. In the models for regions 04 to 07, the fixed effect Mismatch grouping had the values “Match”, “Mismatch” and “Ambiguous”. All six models included the random effect participant. The models were subjected to the `summary` function, and outputs are summarised in Tables 5.9 to 5.11 (some predictor variable names are shortened) and printed in full in Appendix B.1. For reference, the mean reading times are visualized in Figure 5.7. The level of significance (α) was set to the scientific standard of .05. I will only discuss significant effects with a p -value smaller than .05, and I will refrain from discussing non-significant predictor variables unless otherwise stated.

In region 01, the item or proper name region, mean reading time decreased significantly by 28.68 ms for every 1 z increase in `trial_index_z` (standardized measure for sentence pairs read) ($t = -7.17, p < .001$). Also, the item frequency was a significant factor such that for every 1 z increase, participants reading time decreased by 18.75 ms (`item_freq_z`, $t = -4.71, p < .001$).

¹¹Mach and Mismatch values were changed to NonAmbiguous.

In region 02, the item spillover region painted a similar picture. The mean reading time decreased throughout the course of the experiment (`trial_index_z`, $\bar{x} = 28.86$ ms, $t = -12.75$, $p < .001$) and participants read the word 9.95 ms faster for each 1 z increase in `item_freq_z` ($t = -4.43$, $p < .001$). In both regions (01 and 02), ambiguous names were read slower than non-ambiguous names (11.07 ms and 5.42 ms), but the difference was not significant (region 01: $t = 1.13$; region 02 $t = 0.98$).

In region 04, the pronoun region, participants read the pronoun 35.31 ms faster for every 1 z increase in `trial_index_z` ($t = -12.86$, $p < .001$). Further, `participant_mm_grouping` was significant for the difference between the Mismatch condition and the Match condition ($t = 3.35$, $p < .001$). Participants read the pronoun 20.85 ms (Intercept = 519.44) slower when the pronoun mismatched in gender (e. g. *Phillip_{<MALE>}* ... *er_M* ... vs *Phillip_{<MALE>}* ... *sie_F* ...). Another significant influencing factor was `pro` – the pronoun. Participants had slower mean reading times ($\bar{x} = 11.92$ ms) when *sie* was presented compared to the presentation of *er* ($t = 2.18$, $p < .05$).

There was a significant effect for the participant's gender. Individuals who identified as non-binary ($n = 2$) had much faster reading times ($\bar{x} = -210.28$ ms, $t = -2.10$, $p < .05$) than males. The mean reading time difference between females and males was not significant.

In region 05, the first pronoun spillover region, reading times decreased by 28.88 ms for every 1 z increase in `trial_index_z` increase ($t = -17.35$, $p < .001$). As in region 04, the mean reading time in region 05 decreased by 10.00 ms (Intercept = 396.09, `mm_grouping`, Match–Mismatch, $t = 3.78$, $p < .01$), but additionally, there was also an effect for the Ambiguous condition compared to the Match condition (`mm_grouping`, Match–Ambiguous, $t = 4.41$, $p < .05$)). When an ambiguous name was presented at region 01 and any pronoun was presented at region 04 the mean reading time was 9.21 ms slower than the baseline at region 05 (e. g. *Alex_{<MALE/FEMALE>}* ... *er_M/sie_F* *hat* ...). Further, there was an effect of `pro`. Participants read region 05 significantly slower when *sie* was presented compared to *er* independent of matching or mismatching gender (`pro`, $\bar{x} = 7.33$ ms, $t = 2.22$, $p < .05$).

In regions 06 and region 07, the second and third pronoun spillover regions, the only significant effect was the decreased mean reading time throughout the execution of the experiment [region 06: (`trial_index_z`, $\bar{x} = -27.83$ ms, $t = -13.26$, $p < .001$); region 07: (`trial_index_z`, $\bar{x} = -29.24$ ms, $t = -13.26$, $p < .001$)].

Table 5.9: Summary outputs of the Linear Mixed-Effects Models for regions 01 and 02

region 01									
	Categorical Analysis					Continuous Analysis			
Fixed effects:	β	SE	t	p	Fixed effects:	β	SE	t	p
(Intercept)	621.83	76.87	8.09	***	(Intercept)	616.97	77.03	8.01	***
mm_grouping									
Ambiguous	11.07	9.80	1.13		itemPro_mm	2.71	1.58	1.72	.
trial_index_z	-28.68	4.00	-7.17	***	trial_index_z	-29.10	4.00	-7.28	***
list2	-81.75	86.32	-0.95		list2	-81.80	86.41	-0.95	
list3	-3.81	88.52	-0.04		list3	-2.49	88.62	-0.03	
list4	-80.98	87.66	-0.92		list4	-80.68	87.76	-0.92	
list5	29.79	82.91	0.36		list5	30.61	83.00	0.37	
list6	-43.38	86.02	-0.50		list6	-43.99	86.11	-0.51	
pronounSie	-3.15	7.97	-0.40		pronounSie	-3.95	7.96	-0.50	
item_freq_z	-18.75	3.98	-4.71	***	item_freq_z	-19.66	3.99	-4.92	***
genderNA	45.52	95.66	0.48		genderNA	45.61	95.76	0.48	
gendernb	-189.22	147.73	-1.28		gendernb	-187.93	147.89	-1.27	
genderw	12.68	57.53	0.22		genderw	12.69	57.59	0.22	
age_z	-18.12	26.50	-0.68		age_z	-18.09	26.53	-0.68	
region 02									
	Categorical Analysis					Continuous Analysis			
Fixed effects:	β	SE	t	p	Fixed effects:	β	SE	t	p
(Intercept)	454.22	51.52	8.82	***	(Intercept)	450.62	51.52	8.75	***
mm_grouping									
Ambiguous	5.42	5.55	0.98		itemPro_mm	1.69	0.89	1.90	.
trial_index_z	-28.86	2.26	-12.75	***	trial_index_z	-29.04	2.27	-12.84	***
list2	-24.07	57.89	-0.42		list2	-23.97	57.84	-0.41	
list3	-14.37	59.34	-0.24		list3	-14.17	59.30	-0.24	
list4	-29.87	58.77	-0.51		list4	-29.82	58.73	-0.51	
list5	-2.66	55.60	-0.05		list5	-2.47	55.57	-0.04	
list6	2.34	57.68	0.04		list6	2.30	57.64	0.04	
pronounSie	-1.27	4.51	-0.28		pronounSie	-1.48	4.51	-0.33	
item_freq_z	-9.95	2.25	-4.43	***	item_freq_z	-10.37	2.25	-4.60	***
genderNA	65.47	64.16	1.02		genderNA	65.61	64.12	1.02	
gendernb	-140.67	98.98	-1.42		gendernb	-139.92	98.91	-1.42	
genderw	1.34	38.58	0.04		genderw	1.40	38.56	0.04	
age_z	3.36	17.77	0.19		age_z	3.26	17.76	0.18	

Signif. codes: “***” = $p < .001$, “**” = $p < .01$, “*” = $p < .05$, “.” = $p < .1$, “” = $p > .1$

Table 5.10: Summary outputs of the Linear Mixed-Effects Models for regions 04 and 05

region 04									
	Categorical Analysis					Continuous Analysis			
Fixed effects:	β	SE	t	p	Fixed effects:	β	SE	t	p
(Intercept)	519.44	52.10	9.97	***	(Intercept)	518.928	52.09	9.96	***
mm_grouping					itemPro_mm	3.65	1.08	3.37	***
Mismatch	20.85	6.23	3.35	***					
mm_grouping					Ambiguous	9.97	7.29	1.37	
					trial_index_z	-35.31	2.75	-12.86	***
					list2	10.62	58.44	0.18	
					list3	-46.85	59.92	-0.78	
					list4	-53.82	59.34	-0.91	
					list5	17.41	56.13	0.31	
					list6	-12.97	58.23	-0.22	
					pronounSie	11.92	5.47	2.18	*
					item_freq_z	4.77	2.74	1.74	.
					genderNA	50.81	64.77	0.78	
					gendernb	-210.28	100.06	-2.10	*
					genderw	7.63	38.95	0.20	
					age_z	-7.94	17.94	-0.44	
region 05									
	Categorical Analysis					Continuous Analysis			
Fixed effects:	β	SE	t	p	Fixed effects:	β	SE	t	p
(Intercept)	396.09	38.06	10.41	***	(Intercept)	396.53	38.01	10.43	***
mm_grouping					itemPro_mm	1.79	0.66	2.73	**
Mismatch	9.99	3.78	2.65	**					
mm_grouping					Ambiguous	9.21	4.41	2.09	*
					trial_index_z	-28.88	1.67	-17.35	***
					list2	-10.00	42.73	-0.23	
					list3	-32.10	43.80	-0.73	
					list4	-9.99	43.39	-0.23	
					lis t5	1.30	41.04	0.03	
					list6	0.84	42.58	0.02	
					pronounSie	7.33	3.31	2.22	*
					item_freq_z	-0.55	1.66	-0.33	
					genderNA	51.21	47.35	1.08	
					gendernb	-101.11	73.07	-1.38	
					genderw	-0.65	28.48	-0.02	
					age_z	6.28	13.12	0.48	

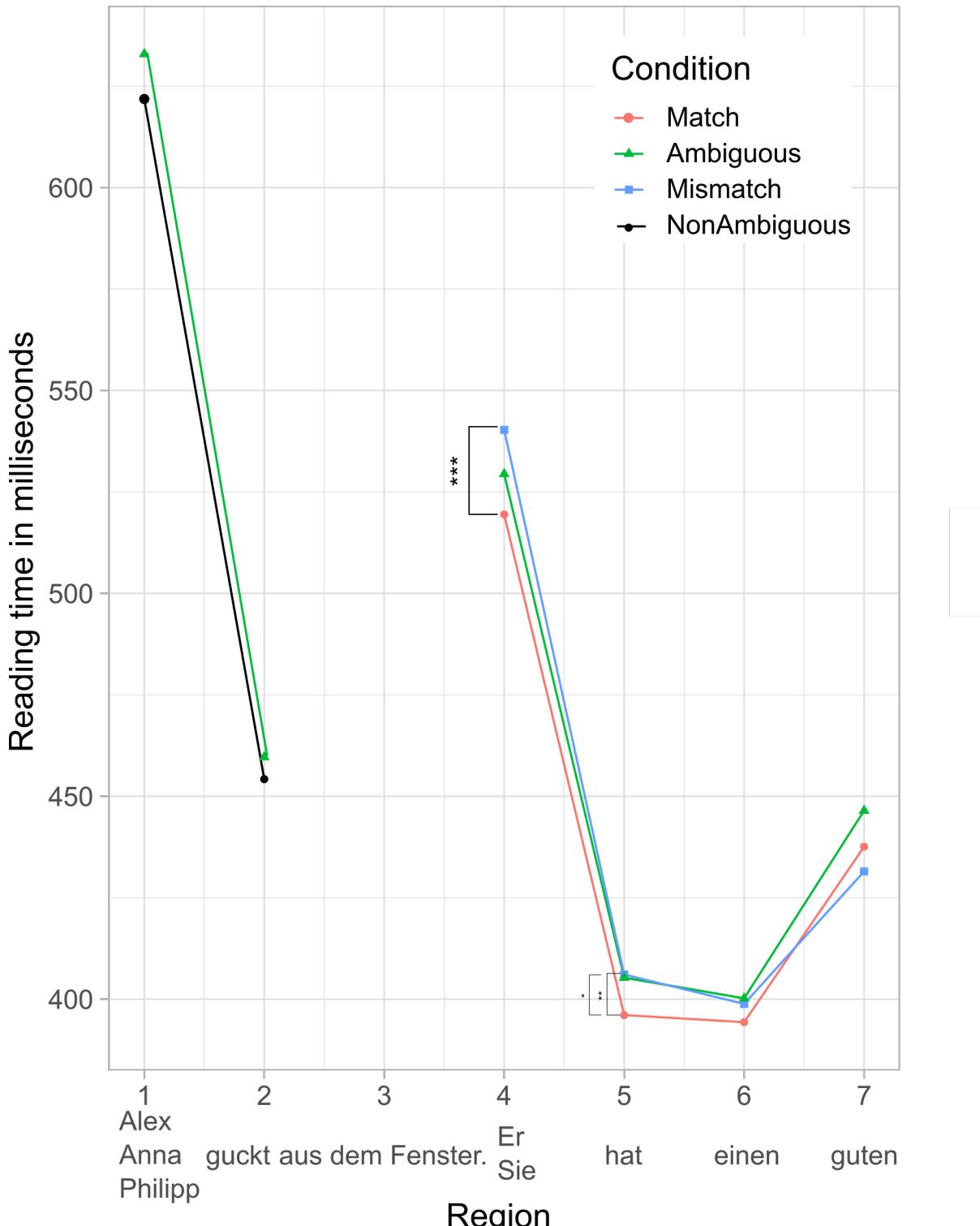
Signif. codes: “***” = $p < .001$, “**” = $p < .01$, “*” = $p < .05$, “.” = $p < .1$, “” = $p > .1$

Table 5.11: Summary outputs of the Linear Mixed-Effects Models for regions 06 and 07

region 06									
	Categorical Analysis					Continuous Analysis			
Fixed effects:	β	SE	t	p	Fixed effects:	β	SE	t	p
(Intercept)	394.33	39.16	10.07	***	(Intercept)	394.23	39.15	10.07	***
mm_grouping					itemPro_mm	1.02	0.67	1.51	
Mismatch	4.50	3.88	1.16						
mm_grouping									
Ambiguous	5.87	4.56	1.29						
trial_index_z	-27.83	1.71	-16.29	***	trial_index_z	-27.89	1.71	-16.34	***
list2	-13.20	43.96	-0.30		list2	-13.04	43.95	-0.30	
list3	-44.68	45.06	-0.99		list3	-44.41	45.05	-0.99	
list4	-16.15	44.63	-0.36		list4	-15.96	44.63	-0.36	
list5	-0.39	42.22	-0.01		list5	-0.23	42.22	-0.01	
list6	-16.88	43.80	-0.39		list6	-16.70	43.79	-0.38	
pronounSie	2.67	3.41	0.78		pronounSie	2.64	3.41	0.77	
item_freq_z	-1.95	1.71	-1.14		item_freq_z	-2.07	1.71	-1.21	
genderNA	71.33	48.70	1.47		genderNA	71.35	48.69	1.47	
gendernb	-101.13	75.16	-1.35		gendernb	-100.95	75.16	-1.34	
genderw	0.59	29.29	0.02		genderw	0.58	29.29	0.02	
age_z	4.18	13.50	0.31		age_z	4.16	13.50	0.31	
region 07									
	Categorical Analysis					Continuous Analysis			
Fixed effects:	β	SE	t	p	Fixed effects:	β	SE	t	p
(Intercept)	437.58	48.33	9.05	***	(Intercept)	439.00	48.13	9.12	***
mm_grouping					itemPro_mm	-0.79	0.87	-0.91	
Mismatch	-6.10	5.01	-1.22						
mm_grouping									
Ambiguous	8.88	5.85	1.52						
trial_index_z	-29.24	2.21	-13.26	***	trial_index_z	-29.47	2.21	-13.36	***
list2	-5.04	54.25	-0.09		list2	-4.48	54.03	-0.08	
list3	-33.98	55.62	-0.61		list3	-32.80	55.38	-0.59	
list4	-15.95	55.09	-0.29		list4	-16.07	54.86	-0.29	
list5	-1.78	52.11	-0.03		list5	-1.18	51.90	-0.02	
list6	10.56	54.06	0.20		list6	11.13	53.83	0.21	
pronounSie	0.42	4.39	0.10		pronounSie	0.07	4.39	0.02	
item_freq_z	-3.57	2.20	-1.62		item_freq_z	-3.79	2.20	-1.72	
genderNA	79.62	60.11	1.32		genderNA	79.99	59.86	1.34	
gendernb	-126.13	92.79	-1.36		gendernb	-125.90	92.40	-1.36	
genderw	3.54	36.16	0.10		genderw	3.69	36.00	0.10	
age_z	10.45	16.66	0.63		age_z	10.28	16.59	0.62	

Signif. codes: “***” = $p < .001$, “**” = $p < .01$, “*” = $p < .05$, “.” = $p < .1$, “” = $p > .1$

Figure 5.7: Reading time for each condition of participant_mm_grouping(_nonAmb) for presentation regions 01 to 07



5.4.2 Continuous Analysis

I repeated the process for the second predictor of theoretical interest, `participant_itemPro_mm_num` (`itemPro_mm_num`) or the mismatch gradient. I fitted a linear mixed-effects regression model for every region using the fixed effects and random effect described above except `participant_mm_grouping_nonAmb/participant_mm_grouping`, which was substituted with `participant_itemPro_mm_num`. The new predictor is a numeric variable depicting the participant's perceived mismatch between the pronoun's gender and the item's gender on a gradient between 0 and 6. For example, a fictional participant encountered the sentence pair with the item and pronoun combination *Kim ... er* The participant rated *Kim* as “neutral”, which equates to the value 4 on a 7-point scale. The (absolute) difference between the item value (4) and the pronoun value (*er* = 7, *sie* = 1) results in the item–pronoun Mismatch value 3.

The results of the numeric predictor are nearly identical to the results of the categorical predictor (cf. Tables 5.9 to 5.11). All significant effects in all regions, including the strength of the effect are identical. I will refrain from repeating myself, so I will only describe the effect of `itemPro_mm` below.

In region 04, participants read the pronoun 3.65 ms slower for every degree of mismatch increase (`participant_itemPro_mm`, $t = 3.37$, $p < .001$). Given 0 is a full match, and 6 is a full mismatch, then for every 1 step increase, participants slowed down by 3.65 ms.

In region 05, as compared to region 04, the effect of `participant_itemPro_mm` persisted but with approximately half the impact. While at region 04, participants had a 3.65 ms longer mean reading time, one word later, they had a 1.79 ms longer mean reading time. The effect remained significant ($t = 2.73$, $p < .01$).

Table 5.12: Table of coefficients of determination for all regions of interest

	Categorical Analysis		Continuous Analysis	
	marginal R^2	conditional R^2	marginal R^2	conditional R^2
region 01	0.06	0.49	0.06	0.49
region 02	0.06	0.57	0.06	0.57
region 04	0.08	0.49	0.08	0.49
region 05	0.09	0.59	0.09	0.59
region 06	0.09	0.59	0.09	0.59
region 07	0.07	0.56	0.07	0.56

I calculated the marginal R^2 and the conditional R^2 for all regions (see Table 5.12). The marginal R^2 expresses how much variance is explained by the fixed effects, while the con-

ditional R^2 also includes the random effect(s). Overall, 54.83% is explained by all predictor variables, but only 7.50% is explained by the marginal R^2 . This means that nearly half (47.33%) of all variance in the data stems from the only random effect – participant.

5.4.3 Comparing Significant Regions

Regions 04 and 05 are significantly affected by the misaligned gender. Figures 5.8a to 5.8d illustrate the distribution of reading time across the two regions and types of classifying the data. Figure 5.8a shows the shortest reading time for the Match condition illustrated by the *mean* and *IQR*, and a shorter reading time in the Ambiguous condition than the Match condition. The conditions in region 05 (Figure 5.8b) follow the same pattern. The distribution of reading time is strongly negatively skewed for all conditions. All show a long tail, whereas the tails are more pronounced in region 04 than in region 05. A unique pattern is a bimodal distribution in the Ambiguous condition region 04 but not in region 05.

Figures 5.8c and 5.8d show the distribution of the mismatch gradient determined by `participant_itemPro_mm_num`. The overall variance across all data points in region 04 is higher compared to region 05, which was already expressed in the long tails above. Most data is found on “0” or “6” on the mismatch gradient, some on “3”, and very little data on the other four spots of the mismatch gradient.

5.4.4 Split Analyses

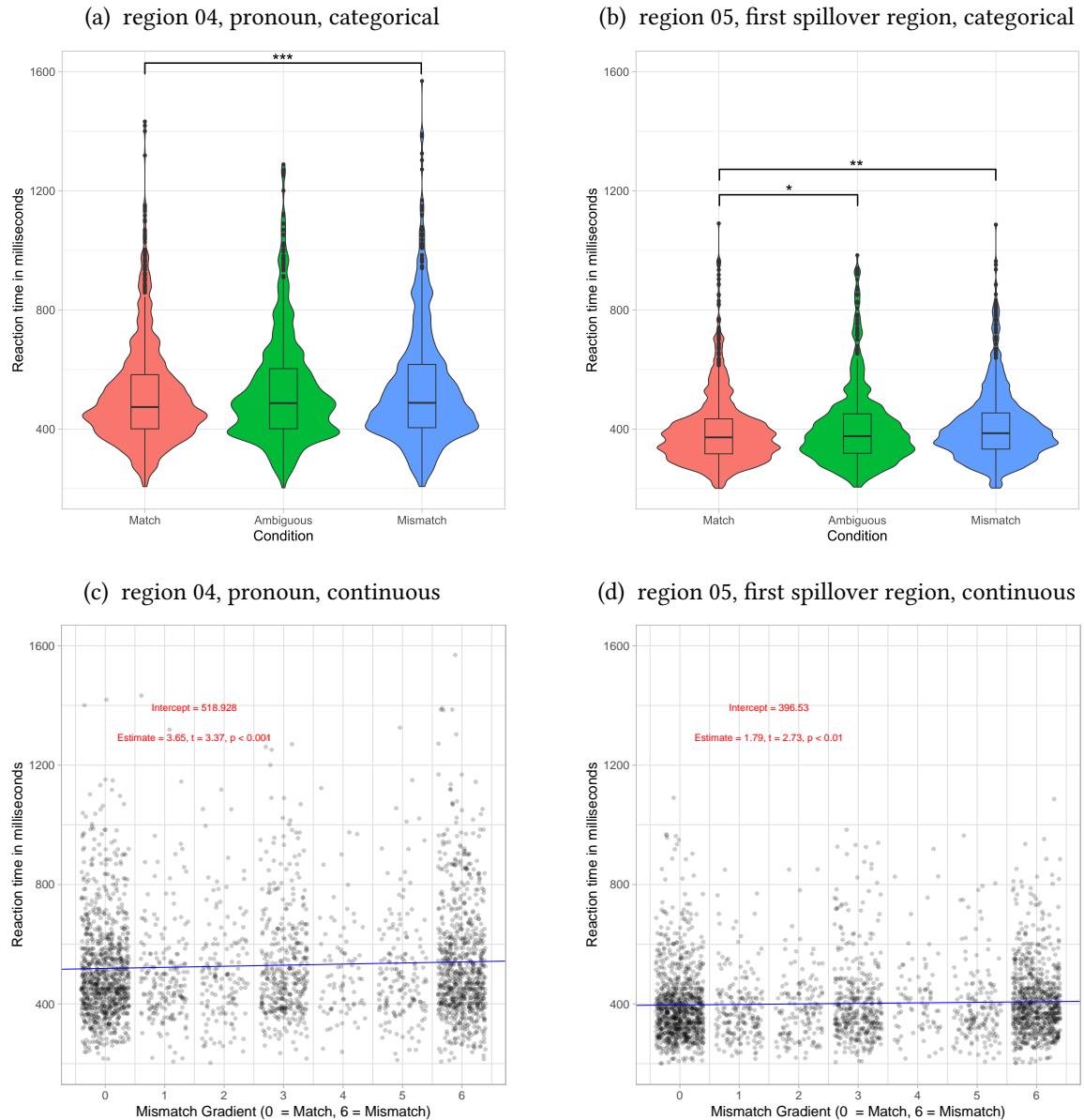
The fixed effect of pronoun (pro) showed a significant effect in region 04 and 05; hence I examined whether the significance of `participant_mm_grouping` persisted given only *er* or *sie* was presented. *Sie* in German is underspecified for 3SG.F, 3PL, and 2SG.FORM. I repeated the analysis with the same fixed effects (except pro) and random effect. I only printed significant effects or non-significant effects, given the other subset shows an effect (see Table 5.13). The full lmer outputs are printed in the Appendices B.3 and B.4.

Region 04 exhibited the same differences in reading time as the full-data frame analysis.

Region 05 showed the only difference between the two subsets and the two regions. In the *er*-subset, the mean reading time was significantly slower in the Mismatch condition than the Match condition ($\bar{x} = 12.91$, $t = 2.38$, $p < .05$). In the *sie*-subset this effect was not significant ($t = 1.06$). The slower reading times of the Ambiguous condition compared to the Match condition seen in the full-data frame vanished in the data frame split. Every other significance (except for pro itself) was identical to the full data frame analysis.

Data frame splits by participant gender and by L2s were conducted but revealed no meaningful differences.

Figure 5.8: Distribution of reading time at region 04 and 05 with participant_mm_grouping (Match, Mismatch, Ambiguous) as the main predictor of interest and reading time increase dependent on the increase in Mismatch value participant_itemPro_mm_num



Signif. codes: “***” = $p < .001$, “**” = $p < .01$, “*” = $p < .05$

Table 5.13: Summary outputs of the *er*-subset and *sie*-subset analyses

region 04								
	sie-subset				er-subset			
Fixed effects:	β	SE	t	p	β	SE	t	p
(Intercept)	538.47	52.30	10.30	***	516.48	51.25	10.08	***
mm_grouping								
Mismatch	18.08	9.01	2.01	*	18.89	8.86	2.13	*
mm_grouping								
Ambiguous	12.55	10.70	1.17		7.67	10.16	0.76	
trial_index_z	-33.09	3.83	-8.63	***	-38.39	4.11	-9.33	***
gendifernb	-214.54	100.71	-2.13	*	-206.23	98.80	-2.09	*
region 05								
	sie-subset				er-subset			
Fixed effects:	β	SE	t	p	β	SE	t	p
(Intercept)	412.09	38.49	10.71	***	389.12	38.59	10.08	***
mm_grouping								
Mismatch	5.64	5.33	1.06		12.91	5.42	2.38	*
mm_grouping								
Ambiguous	10.78	6.36	1.70	.	5.15	6.21	0.83	
trial_index_z	-28.54	2.28	-12.53	***	-31.11	2.51	-12.42	***
region 06								
	sie-subset				er-subset			
Fixed effects:	β	SE	t	p	β	SE	t	p
(Intercept)	399.68	39.16	10.21	***	389.81	39.03	9.99	***
trial_index_z	-28.86	2.17	-13.31	***	-27.15	2.70	-10.08	***
region 07								
	sie-subset				er-subset			
Fixed effects:	β	SE	t	p	β	SE	t	p
(Intercept)	447.32	46.09	9.71	***	50.15	51.86	8.57	***
trial_index_z	-28.99	2.88	-10.08	***	-29.88	3.52	-8.50	***

Signif. codes: “***” = $p < .001$, “**” = $p < .01$, “*” = $p < .05$, “.” = $p < .1$, “” = $p > .1$

Chapter 6

General Discussion

In this section, I will discuss my results in light of previous findings. While doing so, the unmet regression assumptions need to be kept in mind providing healthy scepticism. I will finish with the limitations of this work.

6.1 Discussing the Norming Study

The Norming study has shown that the inter-participant variation for gender ratings is high. The participants could be grouped into (i) those rating gender on a continuum, (ii) those rating gender on a binary scale, and (iii) those rating gender either on the extremes of a scale or in the centre.

The second finding in the Norming study speaks for a difference between role names and proper names. While in Kennison and Trofe, 2003's comprehensive rating study, raters agreed on role names being gender neutral/ambiguous, in my rating study, there was little agreement on a “true” gender-ambiguous name. For example, *student* is gender ambiguous with a *sd* as low as 0.44, but in my study, no first name between mean ratings 3 to 5 was unanimously ambiguous since the *sd* always exceeded 1.00. A fair comparison between these word classes is difficult, but there has been no previous work on referential failure with proper name antecedents. The different rating groups and the high *sd* emphasised the universal need for Post Hoc studies such that “true” ambiguous names are found on a by-participant basis.

6.2 Discussing the Main Study

Hypothesis 1 stated that the mean reading time would be slower in the Mismatch condition compared to the Match condition. The results show that the anaphora’s gender is immediately compared to that of its antecedent. When the gender of the anaphora is incongruent with the gender of its antecedent, the parser immediately updates the situation model, which costs reading time. The effect is noticeable in the presentation region of the pronoun and

the subsequent word, rejecting the null hypothesis of H1a and H1b (regions 04 and 05) while confirming it for H1c and H2d (regions 06 and 07).

This finding is in line with other works on reading time (Carreiras et al., 1996; Irmén et al., 2010; Kennison & Trofe, 2003) and expands the gender mismatch effect from role names to proper names. In comparison to Kennison and Trofe (2003), the effect was found one region earlier since they had significances in the two regions after the pronoun, whereas my finding suggests immediate resolution. I attribute the difference between proper name results and role name results to either (i) the high importance of names in our daily life¹ compared to role names or (ii) the fact that there was no timeout in my experiment, so participants had no need to move to the next region before they understood that the pronoun's gender was mismatching.

The finding is also in line with Irmén and Schumann (2011) and Schmitt et al. (2002), who used Eye-Tracking and EEG to investigate anaphora resolution. Irmén and Schumann (2011: 1012) found an immediate increase in fixation time for mismatching unambiguous stereotypical gender in German. Schmitt et al. (2002) showed ERPs indicating semantic processing and syntactic reanalysis when a (regular) anaphora² mismatched its antecedent in gender.

In sum, mismatching gender effects in anaphora resolution have been shown across a vast spectrum of languages and measurement methods. With the necessary scepticism kept in mind, I demonstrated that the effect previously found with role name antecedents also extends to first name antecedents.

Hypothesis 2 stated that the mean reading time would be slower in the Ambiguous condition compared to the Match condition. The results rejected the null hypothesis for the first spillover region (H2b) and confirmed it for the remaining regions/hypotheses. In line with Irmén et al. (2010) and Irmén and Schumann (2011), this shows for ambiguous referents, there is a late resolution effect.

Irmén et al. (2010)'s results argued for the linguistic claim that German masculine nouns can be generic. The researchers show anaphora resolution is cognitively more difficult due to the underspecified or ambiguous nature of this word class compared to female role names with non-ambiguous gender. The follow-up study with more fine granular Eye-Tracking data (Irmén & Schumann, 2011: 1012) underpinned the finding. Masculine role noun antecedents indicated late processing, while feminine role names indicated early processing of mismatching gender. These effects are mirrored in my results of immediate reading time effects for unambiguous names and a delayed effect for ambiguous names. In contrast, recent research used discriminative learning to show that the semantic vector of the generic masculine is highly similar to the vector of the explicit masculine and dissimilar to the explicit feminine,

¹They are (probably) after pronouns the most frequent referring expression to (other) humans.

²I have highlighted “regular” because Schmitt et al. (2002) used next to *er* and *sie* also *es* as an anaphora which is a-typical in German. The results I am referring to come from the *er/sie*-data.

suggesting that the generic masculine is not gender neutral (Schmitz, 2022; Schmitz et al., 2023), so the generic masculine of role names is put into question.

Bjorkman (2017)'s work on *they* suggested an acceptability scale. This thesis showed at the pronoun region longer mean reading time in the Ambiguous condition (n.s.) of approximately half the size of the Mismatch condition. Further, in the subsequent region, the effect was also smaller in the Ambiguous condition than in the Mismatch condition, so a mismatch scale (*Anna...sie > Alex...sie > Phillip...sie*) analogous to the *they*-acceptability scale (*everyone...they > cousin...they > father...they*) is supported.

The Continuous Analysis, with its numeric predictor variable, showed, on the one hand, that there is a significant increase on a mismatch-gradient basis, but, on the other hand, most data is at the extremes of the continuous gender scale. This is in line with Ackerman (2019)'s exemplar tier, which also has most exemplars at the extremes on the continuum.

Figures 5.8c and 5.8d showed that many participants rated names as “neutral/ambiguous”. This mostly fits into Ackerman, 2019's description of a one-to-one but also a one-to-many mapping of lemmas and grammatical gender. Maybe there is also a one-to-many mapping of stereotypical gender so that *Alex* is mapped to the stereotypical gender <MALE> and <FEMALE>. Cacciari et al. (1997) believes that there are role names with two (stereotypical) genders, so a double assignment of gender onto first names is feasible.

Another relatively straightforward approach is multiple name activation which activates, as a logical consequence, multiple genders. Valentine et al. (1996) mentions competition among the names during name recognition but does not exclude multiple Person Identity Node activations. My research cannot validate either one-to-many gender mapping or multiple activations, but the lack of data at mismatch gradient values 1, 2, 4, and 5 in Figures 5.8c and 5.8d indicates that gender is not seen as an evenly distributed continuum or at least it is not rated as such.

Split and Pronoun Analysis shows that the slight bimodal distribution in the Ambiguous condition at the pronoun region (see Figure 5.8a) could not be explained by any data frame split. The effect of a longer mean reading time in region 05 vanished in the pronoun split analysis. I assume that halving the data – decreasing the statistical power – is the cause of the effect's absence.

The significant difference between Mismatch and Match in region 05 in the *er*-subset but not in the *sie*-subset is surprising and could also be attributed to the loss in statistical power. I would have expected *sie* to have the longer-lasting effect due to the syncretic nature of the pronoun. 3SG.F, 3PL, and 2SG.FORM are form identical incorporated in *sie*, and I expect *sie* to cause greater cognitive effort.

In the main analysis, *sie* was read slower overall. Kennison and Trofe (2003), who found a main effect for pronoun too, attributed the effect to printed frequency but in German *sie* has

a higher printed frequency than *er* based on the *DWDS-Zeitungskorpus (ab 1945)* (DWDS – Digitales Wörterbuch der deutschen Sprache, n.d.). I attribute the effect to the aforementioned syncretism.

Hypothesis 3 stated the mean reading time during initial gender encoding would be significantly longer for ambiguous names than for non-ambiguous names. For H3a and H3b, the null hypothesis was confirmed since the difference shown was not significant. Either encoding two genders is no more difficult than encoding one, or deciding on one does not lead to greater cognitive effort.

6.3 Criticism

Common sense says one cannot treat ordinal numbers like continuous numbers. What is the difference between “sehr männlich” and an unlabeled point on a 7-point Likert scale, and is the difference between two points the same as the difference between the next two points? Probably not. Still, many researchers (Carreiras et al., 1996; Duffy & Keir, 2004; Kennison & Trofe, 2003; Osterhout, 1997; Osterhout & Mobley, 1995; Shinar, 1975; Valentine et al., 1991) treated ordinal scales as if they were continuous, and I did too. To be in line with this common research practice does not resolve the problem, but Williams (2020) states that using ordinal independent variables as predictor variables is statistically acceptable.

The model criticism showed that all three regression assumptions are unmet, such that general research practice deems the linear mixed-effects models unacceptable. Schielzeth et al. (2020) found evidence that `lmer` models are very robust towards even severe model assumption violations. Substantially skewed and heteroscedastic distributions could yield overall good results, so fulfilling the assumptions is good, but unmet assumptions do not need to be a knock-out criterion for statistical evaluation.

In future work, I would improve these experiments in various ways. There was no time-out in the self-paced reading experiment, which led to extreme outliers of up to one minute “reading time” for one word. I removed the whole sentence in which reading times of over 2 seconds occurred, leading to the immense data loss of 27 % (before the residual trim). Further, the filler items (role names) were randomly distributed over the carrier sentences leading to semantically irritating NP and PP combinations (see Example (1)). The example also illustrates that not all carrier sentences are stereotypically neutral. The full list found in Appendix A shows that most carrier sentences provide a neutral context. Also, target items and filler items showed gender incongruity such that the topic of gender was very present for the participants. A better variability within the filler items could have helped the study.

- (1) *Die Flugbegleiterin renoviert in der Garage. Er möchte die neuen Werkzeuge testen.*
(*The flight attendant is renovating in the garage. He wants to test the new tools.*)

An additional concern, next to the inter-participant variation caught with the Post Hoc study, is the intra-participant variation. It is unclear whether the name–gender attribution during the Main study is the same attribution as during the Post Hoc study since there was a two-week delay between the two experiments. For example, just before the Post Hoc study, a participant received a message from a female friend called *Charlie*, who is generally not prominent in the participant’s mental lexicon (due to little contact). This interaction could have primed the name–gender association such that the belief system during the Main study and the Post Hoc study is not the same. On the other hand, if the studies were conducted as one, the Main study might have primed the Post Hoc study ratings. A smaller time window could have improved the study design.

For one of the biggest points of critique, I thank a listener of my talk. He pointed out that there is not only the name frequency that I measured with Google queries but also the individual name frequency for every participant. To exemplify: The infrequent name *Mathilda* could have a high individual frequency because *Mathilda* is a participant’s mother’s name. The random effect participant explained a lot of variance in the data (approximately 40 %), but the additional variable would have improved the model and should be good practice for future Post Hoc studies. Additionally, I did not include word length in the model, which is critical for reading time. These factors combined should improve the marginal R^2 of 7.50%.

Chapter 7

Conclusion

This thesis presents novel findings which question the validity of some stimuli used in experimental research on anaphora resolution. With the statistical and methodological limitations in mind, I showed that referential failure effects extend from role names to first names, and the effect – a longer reading time – is also present for ambiguous first name antecedents independent of the pronoun. In the sentence pair, *Alex liest diese Thesis. Er denkt, dass das Thema relevant ist.* the first word after the pronoun (*denkt*) is read slower¹, regardless of whether *er* or *sie* refers to *Alex*. It remains unclear whether the pronouns always mismatch because the effect could also originate from an initial underspecification and a subsequent gender assignment at the pronoun. *Anna* or *Phillip*, as antecedents, show in the Mismatch condition, compared to the Match condition, an immediate reading time penalty at the pronoun and the following word. My results mirror findings from high-resolution Eye-Tracking studies in German (Irmens & Schumann, 2011). They argue “statistical analyses should always include grammatical gender as an experimental factor rather than collapsing across masculine and feminine forms and contrasting gender-congruent and incongruent experimental conditions” (Irmens & Schumann, 2011: 1012). Findings from my Norming study extend this recommendation.

In the rating study, name–gender associations show high variation across participants – especially for ambiguous names². This is, on the one hand, evidence that role names and proper names are different (cf. Valentine et al., 1996), and on the other hand, the results show the importance of capturing name–gender associations on a subjective by-participant level. As an extension of Kennison and Trofe (2003), the Post Hoc study provided valuable data, which allowed for the statistical analysis that sees gender mismatch as a continuum and extrapolates gender mismatch categories/conditions from the continuum. This multi-level approach closely resembles Ackerman (2019)’s gender framework, which is, in turn, based on exemplar and prototype theory. The analysis of gender mismatch as a continuum is significant but, on close inspection, is not rated as one. In support of Irmens and Schumann

¹I compared the Ambiguous condition (e.g. *Alex ... er*) with the Match condition (e.g. *Phillip ... er*).

²Role name ratings showed agreement among the “neutral” cluster (e.g. *student*) (Kennison & Trofe, 2003).

(2011), I recommend measuring the participant's unique understanding of gender if gender has a prominent role in the given research.

Another interesting observation is that the two non-binary participants seemingly ignored the mismatching pronouns. If the effect was a result of the small sample size or an effect stemming from a non-binary gender distribution in non-binary minds remains to be investigated in future research.

I am excited to see researchers use the normed material found in the appendix or use the new method of analyzing gender mismatch. The next steps towards a better understanding of ambiguous names as a factor in psycholinguistic research could entail studies utilizing haemodynamic methods providing better resolution and the interpretation of polarity, and the incorporation of large language models.

Bibliography

- Ackerman, L. (2019). Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1), 1–27. <https://doi.org/10.5334/gjgl.721>
- Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16(5), 443–463. [https://doi.org/10.1016/0378-2166\(91\)90136-L](https://doi.org/10.1016/0378-2166(91)90136-L)
- Baayen, R. H. (2011). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bjorkman, B. M. (2017). Singular *they* and the syntactic representation of gender in English. *Glossa: a journal of general linguistics*, 2(1), 1–13. <https://doi.org/10.5334/gjgl.374>
- Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of Memory and Language*, 125, 1–13. <https://doi.org/10.1016/j.jml.2022.104334>
- Cacciari, C., Carreiras, M., & Cionini, C. B. (1997). When Words Have Two Genders: Anaphor Resolution for Italian Functionally Ambiguous Words. *Journal of Memory and Language*, 37(4), 517–532. <https://doi.org/10.1006/jmla.1997.2528>
- Callahan, S. M. (2008). Processing anaphoric constructions: Insights from electrophysiological studies. *Journal of Neurolinguistics*, 21(3), 231–266. <https://doi.org/10.1016/j.jneuroling.2007.10.002>
- Cao, Y. T., & Daumé, H. (2021). Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle. *Computational Linguistics*, 47(3), 615–661. https://doi.org/10.1162/coli_a_00413
- Carreiras, M., Garnham, A., & Oakhill, J. (1993). The use of superficial and meaning-based representations in interpreting pronouns: Evidence from Spanish. *European Journal of Cognitive Psychology*, 5(1), 93–116. <https://doi.org/10.1080/09541449308406516>
- Carreiras, M., Garnham, A., Oakhill, J., & Cain, K. (1996). The Use of Stereotypical Gender Information in Constructing a Mental Model: Evidence from English and Spanish. *The Quarterly Journal of Experimental Psychology Section A*, 49(3), 639–663. <https://doi.org/10.1080/713755647>

- Chow, W.-Y., Lewis, S., & Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution [Publisher: Frontiers Media SA]. *Frontiers in Psychology*, 5, 630. <https://doi.org/10.3389/fpsyg.2014.00630>
- Cohen, G., & Burke, D. M. (1993). Memory for proper names: A review. *Memory*, 1(4), 249–263. <https://doi.org/10.1080/09658219308258237>
- Dell, G. S., & O’Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42(1-3), 287–314. [https://doi.org/10.1016/0010-0277\(92\)90046-K](https://doi.org/10.1016/0010-0277(92)90046-K)
- Deutschlernerblog. (2019). 1000 Orte mit Präpositionen - Ortsangaben: Welche Präposition? Retrieved January 13, 2023, from <https://deutschlernerblog.de/1000-orte-mit-praepositionen-lokale-praepositionen-welche-praeposition/>
- Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, 32(4), 551–559. <https://doi.org/10.3758/bf03195846>
- DWDS – Digitales Wörterbuch der deutschen Sprache. (n.d.). DWDS-Wortverlaufskurve für „er · sie“, erstellt durch das Digitale Wörterbuch der deutschen Sprache. <https://www.dwds.de/r/plot/?view=1&corpus=zeitungenxl&norm=date%2Bclass&smooth=spline&genres=0&grand=1&slice=1&prune=0&window=3&wbase=0&logavg=0&logscale=0&xrange=1946%3A2022&q1=er&q2=sie>
- Frege, G. (1948). Sense and Reference. *The Philosophical Review*, 57(3), 209–230. <https://doi.org/10.2307/2181485>
- Garnham, A., Oakhill, J., & Cruttenden, H. (1992). The role of implicit causality and gender cue in the interpretation of pronouns. *Language and Cognitive Processes*, 7(3-4), 231–255. <https://doi.org/10.1080/01690969208409386>
- Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17(3), 311–347. <https://www.sciencedirect.com/science/article/pii/S0364021305800023>
- Grosz, B. J., Joshi, A. K., & Weinstein, S. (1986). Towards a computational theory of discourse interpretation. *Unpublished manuscript*.
- Hammer, A., Jansma, B. M., Lamers, M., & Münte, T. F. (2005). Pronominal Reference in Sentences about Persons or Things: An Electrophysiological Approach. *Journal of Cognitive Neuroscience*, 17(2), 227–239. <https://doi.org/10.1162/0898929053124947>
- Irmén, L. (2007). What’s in a (Role) Name? Formal and Conceptual Aspects of Comprehending Personal Nouns. *Journal of Psycholinguistic Research*, 36(6), 431–56. <https://doi.org/10.1007/s10936-007-9053-z>
- Irmén, L., Holt, D. V., & Weisbrod, M. (2010). Effects of role typicality on processing person information in German: Evidence from an ERP study. *Brain Research*, 1353, 133–144. <https://doi.org/10.1016/j.brainres.2010.07.018>

- Irmen, L., & Kurovskaja, J. (2010). On the Semantic Content of Grammatical Gender and Its Impact on the Representation of Human Referents. *Experimental Psychology*, 57(5), 367–375. <https://doi.org/10.1027/1618-3169/a000044>
- Irmen, L., & Schumann, E. (2011). Processing grammatical gender of role nouns: Further evidence from eye movements. *Journal of Cognitive Psychology*, 23(8), 998–1014. <https://doi.org/10.1080/20445911.2011.596824>
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329–354. <https://doi.org/10.1037/0033-295X.87.4.329>
- Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and Coreference Revisited. *Journal of Semantics*, 25(1), 1–44. <https://doi.org/10.1093/jos/ffm018>
- Kennison, S. M., & Trofe, J. L. (2003). Comprehending Pronouns: A Role for Word-Specific Gender Stereotype Information. *Journal of Psycholinguistic Research*, 32(3), 355–378. <https://doi.org/10.1023/A:1023599719948>
- Kreiner, H., Garrod, S. C., & Sturt, P. (2013). Number agreement in sentence comprehension: The relationship between grammatical and conceptual factors. *Language and Cognitive Processes*, 28(6), 829–874. <https://doi.org/10.1080/01690965.2012.667567>
- Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(01). <https://doi.org/10.1017/S0140525X99001776>
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
- McKoon, G., Greene, S. B., & Ratcliff, R. (1993). Discourse models, pronoun resolution, and the implicit causality of verbs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(5), 1040–1052. <https://doi.org/10.1037/0278-7393.19.5.1040>
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207–238. <https://doi.org/10.1037/0033-295X.85.3.207>
- Nieuwland, M. S., Petersson, K. M., & van Berkum, J. J. A. (2007). On sense and reference: Examining the functional neuroanatomy of referential processing. *NeuroImage*, 37(3), 993–1004. <https://doi.org/10.1016/j.neuroimage.2007.05.048>
- Osterhout, L. (1997). On the Brain Response to Syntactic Anomalies: Manipulations of Word Position and Word Class Reveal Individual Differences. *Brain and Language*, 59(3), 494–522. <https://doi.org/10.1006/brln.1997.1793>
- Osterhout, L., & Mobley, L. A. (1995). Event-Related Brain Potentials Elicited by Failure to Agree. *Journal of Memory and Language*, 34(6), 739–773. <https://doi.org/10.1006/jmla.1995.1033>
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>

- Peterson, R., A. (2021). Finding Optimal Normalizing Transformations via bestNormalize. *The R Journal*, 13(1), 310–329. <https://doi.org/10.32614/RJ-2021-041>
- Peterson, R. A., & Cavanaugh, J. E. (2020). Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, 47(13-15), 2312–2327. <https://doi.org/10.1080/02664763.2019.1630372>
- R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573–605. [https://doi.org/10.1016/0010-0285\(75\)90024-9](https://doi.org/10.1016/0010-0285(75)90024-9)
- Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions (C. Sutherland, Ed.). *Methods in Ecology and Evolution*, 11(9), 1141–1152. <https://doi.org/10.1111/2041-210X.13434>
- Schmitt, B. M., Lamers, M., & Münte, T. F. (2002). Electrophysiological estimates of biological and syntactic gender violation during pronoun processing. *Cognitive Brain Research*, 14(3), 333–346. [https://doi.org/10.1016/S0926-6410\(02\)00136-2](https://doi.org/10.1016/S0926-6410(02)00136-2)
- Schmitz, D. (2022). In German, all professors are male. <https://doi.org/10.31234/osf.io/yjuhc>
- Schmitz, D., Schneider, V., & Esser, J. (2023). No genericity in sight: An exploration of the semantics of masculine generics in German. <https://doi.org/10.31234/osf.io/c27r9>
- Semenza, C., & Zettin, M. (1988). Generating proper names: A case of selective inability. *Cognitive Neuropsychology*, 5(6), 711–721. <https://doi.org/10.1080/02643298808253279>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591–611. <https://doi.org/10.1093/biomet/52.3-4.591>
- Shinar, E. H. (1975). Sexual stereotypes of occupations. *Journal of Vocational Behavior*, 7(1), 99–111. [https://doi.org/10.1016/0001-8791\(75\)90037-8](https://doi.org/10.1016/0001-8791(75)90037-8)
- Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48(3), 542–562. [https://doi.org/10.1016/S0749-596X\(02\)00536-3](https://doi.org/10.1016/S0749-596X(02)00536-3)
- Swaab, T. Y., Camblin, C. C., & Gordon, P. C. (2004). Electrophysiological Evidence for Reversed Lexical Repetition Effects in Language Processing. *Journal of Cognitive Neuroscience*, 16(5), 715–726. <https://doi.org/10.1162/089892904970744>
- The pandas development team. (2020). Pandas-dev/pandas: Pandas. <https://doi.org/10.5281/zenodo.3509134>
- Valentine, T., Bredart, S., Lawson, R., & Ward, G. (1991). What's in a name? access to information from people's names. *European Journal of Cognitive Psychology*, 3(1), 147–176. <https://doi.org/10.1080/09541449108406224>

- Valentine, T., Brennen, T., & Brédart, S. (1996). *The cognitive psychology of proper names: On the importance of being Ernest*. Routledge.
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.
- van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.
- Williams, R. (2020). Ordinal Independent Variables. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. Williams (Eds.), *SAGE Research Methods Foundations*. SAGE Publications Ltd. <https://doi.org/10.4135/9781526421036938055>
- Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. Routledge. <https://doi.org/10.4324/9781315165547>

Data and Code Availability

The code, the experiment files, and much of the data that was used to obtain the findings of these studies are open-source on OSF at [DOI 10.17605/OSF.IO/JGZQK](https://doi.org/10.17605/OSF.IO/JGZQK) and on GitHub at <https://github.com/alexanderclemen/AmbiguousNamesReferentialFailure> under the [GNU Affero General Public License v3.0](#).

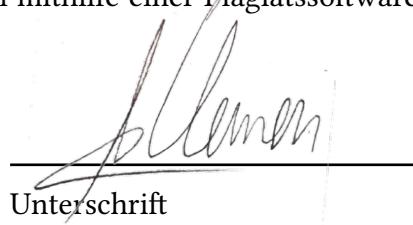
Word Count: 12916

Eigenständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem einzelnen Fall unter Angabe der Quelle kenntlich gemacht. Dies gilt auch für verwendete Zeichnungen, Skizzen, Ton- und Videoaufnahmen sowie graphische Darstellungen. Ich erkläre mich damit einverstanden, dass meine Arbeit im Verdachtsfall mithilfe einer Plagiatssoftware überprüft wird.

Düsseldorf, 20.05.2023

Ort, Datum



Unterschrift

Appendix A

Materials

List of Carrier Sentences

- [Item] spaziert ins Bistro. [Pronoun] möchte die volle Treuekarte einlösen.
- [Item] schreit in der Sauna. [Pronoun] hat einen heißen Aufgussstein berührt.
- [Item] starrt auf die Speisekarte. [Pronoun] möchte die lokalen Köstlichkeiten ausprobieren.
- [Item] fällt aus dem Bett. [Pronoun] hat einen schlimmen Albtraum gehabt.
- [Item] reist in die Metropole. [Pronoun] möchte die weltbekannte Clubkultur erleben.
- [Item] guckt auf den Fahrplan. [Pronoun] hat die heutige Verbindung vergessen.
- [Item] geht zur Pommesbude. [Pronoun] hat die grauenvolle Abnehmkur überstanden.
- [Item] parkt auf dem Radweg. [Pronoun] möchte ein starkes Zeichen setzen.
- [Item] flitzt aus der Behörde. [Pronoun] muss den letzten Bus bekommen.
- [Item] steigt aus dem Zug. [Pronoun] hat das graue Hemd durchgeschwitzt.
- [Item] flüchtet aus dem Restaurant. [Pronoun] hat die hohe Preise unterschätzt.
- [Item] weint zu Hause. [Pronoun] hat mit den Geschwistern Streit.
- [Item] flieht aus dem Fahrstuhl. [Pronoun] hat eine riesige Spinne gesehen.
- [Item] reist zum Turnier. [Pronoun] hat das ganze Jahr trainiert.
- [Item] strickt im Pflegeheim. [Pronoun] hat eine gute Freundschaft geschlossen.
- [Item] jongliert im Freizeitpark. [Pronoun] hat einen neuen Job gefunden.
- [Item] liegt im Liegestuhl. [Pronoun] hat eine missglückte Knie-OP erlitten.
- [Item] hüpfst auf dem Trampolin. [Pronoun] möchte die neuen Nachbarskinder bespäßen.
- [Item] erwacht von der Weinprobe. [Pronoun] hatte einen spaßigen Abend genossen.
- [Item] reitet aus dem Stall. [Pronoun] hat die langweiligen Probestunden absolviert.
- [Item] joggt im Park. [Pronoun] möchte den winterlichen Bauchspeck loswerden.
- [Item] fällt auf der Beerdigung. [Pronoun] hat das tiefe Loch übersehen.
- [Item] starrt auf den Schulhof. [Pronoun] hat einen potenziellen Profispieler gefunden.
- [Item] hüpfst in der Küche. [Pronoun] möchte den oberen Hängeschrank erreichen.
- [Item] schwimmt in der Ostsee. [Pronoun] hat das kalte Wasser gern.
- [Item] erwacht in der Einfahrt. [Pronoun] hat den einzigen Haustürschlüssel verloren.
- [Item] landet in der Notaufnahme. [Pronoun] hat die schweren Handwerksarbeiten unterschätzt.
- [Item] posiert auf dem Plakat. [Pronoun] hat einen tollen Werbedeal bekommen.
- [Item] springt vom Beckenrand. [Pronoun] möchte den schönen Bademeister beeindrucken.
- [Item] kehrt im Stall. [Pronoun] muss die aufgetragenen Sozialstunden abarbeiten.

[Item] posiert am Klavier. [Pronoun] möchte das große Publikum beeindrucken.
[Item] kommt vom Kongress. [Pronoun] hat die alljährliche Zusammenkunft genossen.
[Item] tanzt auf der Veranstaltung. [Pronoun] hat eine freundliche Tanzgruppe gefunden.
[Item] schwimmt im Zoo. [Pronoun] möchte den jungen Orca retten.
[Item] liegt vor dem Fernseher. [Pronoun] hat ein neues Trainingsprogramm angefangen.
[Item] tüftelt am Schließfach. [Pronoun] hat die wichtige Zahlenkombination vergessen.
[Item] joggt zum PKW. [Pronoun] hat einen wichtigen Termin vergessen.
[Item] schwimmt zum Boot. [Pronoun] möchte die einsame Insel verlassen.
[Item] stolpert aus der Kneipe. [Pronoun] hat das neue Craftbier genossen.
[Item] klettert in der Kletterhalle. [Pronoun] möchte einen sexy Sommerbody bekommen.
[Item] rennt zum Briefkasten. [Pronoun] hat den hübschen Postboten gesehen.
[Item] springt in den Pool. [Pronoun] hat ein ertrinkendes Kind gesichtet.
[Item] kommt aus der Kita. [Pronoun] hat die beiden Zwillinge dabei.
[Item] faulenzt im Sessel. [Pronoun] hat einen harten Arbeitstag gehabt.
[Item] erwacht am Bahnhof. [Pronoun] ist mit dem Nachtzug gefahren.
[Item] kommt von der Toilette. [Pronoun] hat die aktuelle Zeitung ausgelesen.
[Item] klettert vom Balkon. [Pronoun] hat die teure Vase zerdeppert.
[Item] schläft im Betrieb. [Pronoun] möchte das große Projekt beenden.
[Item] eilt auf das Amt. [Pronoun] hatte eine essenzielle Anlage vergessen.
[Item] schleicht ins Haus. [Pronoun] möchte die schlafenden Nachbarn nicht wecken.
[Item] stolpert in die Bar. [Pronoun] hat die erste Anzahlung erhalten.
[Item] flüchtet von der Baustelle. [Pronoun] hat ein wichtiges Warnschild übersehen.
[Item] kommt vom Kiosk. [Pronoun] hat ein leckeres Snickers gekauft.
[Item] marschiert aus dem Rathaus. [Pronoun] hat das goldene Buch beschmutzt.
[Item] sitzt beim Abendessen. [Pronoun] muss die immergleichen Diskussionen ertragen.
[Item] kriecht ins Bad. [Pronoun] hat ein leckeres Bier getrunken.
[Item] kommt von der Bandprobe. [Pronoun] hat ein exzellentes Solo hingelegt.
[Item] kommt vom Klo. [Pronoun] hat die wertvolle Arbeitszeit abgesessen.
[Item] zeichnet in der Vorstadt. [Pronoun] hat ein schönes Model gefunden.
[Item] steigt von der Tribüne. [Pronoun] hat einen ehrenvollen Orden erhalten.
[Item] fliegt auf die Malediven. [Pronoun] hat einen schönen Freundschaft gebucht.
[Item] kniet in der Moschee. [Pronoun] wird das übliche Gebet halten.
[Item] reist ins Bistum. [Pronoun] hat den edlen Bischof vermisst.
[Item] renoviert in der Garage. [Pronoun] möchte die neuen Werkzeuge testen.
[Item] faulenzt im Café. [Pronoun] hat einen städtischen Netzausfall erlitten.
[Item] liegt in der Gasse. [Pronoun] hat die falsche Person angestarrt.
[Item] steigt auf das Skateboard. [Pronoun] möchte die junge Nachbarin beeindrucken.
[Item] strickt auf der Karnevalssitzung. [Pronoun] hat die immergleichen Witze satt.
[Item] schleicht zum Deutschkurs. [Pronoun] hat nur wenig Spaß am Lernen.
[Item] fällt von der Leiter. [Pronoun] hat die oberste Stufe verfehlt.
[Item] schläft auf der Arbeit. [Pronoun] muss die lange Nacht überstehen.
[Item] raucht vor dem Zeitungsstand. [Pronoun] hat die leckere Zigarette verdient.
[Item] steigt auf den Tisch. [Pronoun] hat ein großes Maß geleert.
[Item] landet auf der Titelseite. [Pronoun] hat eine schlimme Tat begangen.
[Item] tüftelt am Fahrrad. [Pronoun] hat einen großen Bolzenschneider gekauft.
[Item] betet auf der Fähre. [Pronoun] hat das andauernde Schaukeln satt.
[Item] stürzt auf dem Radrennen. [Pronoun] hat einen ekstatischen Fan übersehen.

[Item] zeichnet im Bus. [Pronoun] hat ein neues Hobby begonnen.
[Item] segelt in der Bucht. [Pronoun] hat ein gebrauchtes Boot gekauft.
[Item] fliegt aus der Mannschaft. [Pronoun] hat den strengen Schiedsrichter angespuckt.
[Item] rennt in den Laden. [Pronoun] hat einen gruseligen Mann gesehen.
[Item] verzweifelt im Konsulat. [Pronoun] hat den wichtigen Reisepass verlegt.
[Item] läuft zur Meisterschaft. [Pronoun] hat den letzten Bus verpasst.
[Item] kriecht in der Werkstatt. [Pronoun] hat die starke Brille verloren.
[Item] fällt aus dem Rollstuhl. [Pronoun] hat den offenen Gully übersehen.
[Item] verzweifelt im Parkhaus. [Pronoun] hat den letzten Parkplatz übersehen.
[Item] steht in der Raucherecke. [Pronoun] muss die neuen Klassenkameraden beeindrucken.
[Item] wandert vom Berg. [Pronoun] hat die weite Aussicht genossen.
[Item] jubelt auf dem Flohmarkt. [Pronoun] hat eine wertvolle Rarität ersteigert.
[Item] spaziert in die Kneipe. [Pronoun] hat eine saftige Gehaltserhöhung erhalten.
[Item] bangt in der Universität. [Pronoun] hat die wichtige Präsentation vermasselt.
[Item] läuft zur Bäckerei. [Pronoun] hat den notwendigen Kuchen vergessen.
[Item] stürzt von der Bühne. [Pronoun] hat eine lockere Stufe übersehen.
[Item] stürzt im Hallenbad. [Pronoun] hat das Laufen-Verboten Schild ignoriert.
[Item] kommt in den Altbau. [Pronoun] hat eine wichtige Wohnungsbesichtigung vereinbart.
[Item] geht aus dem Theaterstück. [Pronoun] hat eine neue Passion entdeckt.
[Item] rodet vom Hügel. [Pronoun] hat diesen weißen Winter Spaß.
[Item] erwacht in der Villa. [Pronoun] hat einen ausgelassenen Abend gehabt.
[Item] wandert aus der Burg. [Pronoun] hat eine hölzerne Schwert gekauft.
[Item] wartet vor dem Computer. [Pronoun] hat einen langwierigen Rechenprozess gestartet.
[Item] flüchtet in die Besprechung. [Pronoun] hat die endlosen Streitigkeiten satt.
[Item] flieht in die Bibliothek. [Pronoun] möchte die lauten Kollegen nicht hören.
[Item] steht vor LIDL. [Pronoun] muss die wertvollen Pfandflaschen wegbringen.
[Item] tanzt in der Disko. [Pronoun] ist der absolute Mittelpunkt des Abends.
[Item] fällt vom Schemel. [Pronoun] hat die anstrengende Beschäftigung unterschätzt.
[Item] eilt auf den Landsitz. [Pronoun] hat den harten Corona-Maßnahmen vernommen.
[Item] spaziert in die Druckerei. [Pronoun] möchte die unschönen Passbilder abholen.
[Item] landet in der Anstalt. [Pronoun] hat einen schweren Burnout erlitten.
[Item] joggt vor der Ampel. [Pronoun] muss auf das Ampelmännchen warten.
[Item] stürzt beim Marathon. [Pronoun] hat die sportlichen Grenzen erreicht.
[Item] rennt zum Unfallort. [Pronoun] hat die notwendigen Verbände dabei.
[Item] simst im Hörsaal. [Pronoun] findet die andauernde Vorlesung langweilig.
[Item] kommt aus dem Verhör. [Pronoun] hat eine leckere Schokotafel geklaut.
[Item] raucht im U-Bahnhof. [Pronoun] möchte die harten Gesetze missachten.
[Item] spaziert zum Trödelmarkt. [Pronoun] möchte das alte Geschirr ersetzen.
[Item] wartet vor der Kasse. [Pronoun] hat die falsche Schlange gewählt.
[Item] kommt vom Vortrag. [Pronoun] hat heute wieder Nichts gelernt.
[Item] guckt aus dem Fenster. [Pronoun] hat einen guten Freund gesehen.
[Item] fliegt aus der Talkshow. [Pronoun] hat die top-secret Geheimnisse verraten.
[Item] schleicht in den Palast. [Pronoun] möchte das teure Porzellan stehlen.

Table A.1: List of rated names in ascending mean gender rating order

Name	mean (sd)	\tilde{x}	Name	mean (sd)	\tilde{x}	Name	mean (sd)	\tilde{x}
Jakob	1.06 (0.34)	1	Noah	1.86 (1.12)	1	Fenja	6.29 (1.02)	7
Georg	1.09 (0.37)	1	Gabriel	1.86 (1.38)	1	Thea	6.34 (1.19)	7
Julius	1.09 (0.37)	1	Dylan	1.97 (1.22)	1	Wiebke	6.37 (1.31)	7
Moritz	1.11 (0.32)	1	Kai	2.14 (1.46)	1	Lia	6.43 (0.95)	7
Paul	1.11 (0.32)	1	Chris	2.17 (1.29)	2	Maria	6.54 (0.78)	7
Tobias	1.11 (0.32)	1	Liam	2.17 (1.48)	2	Merle	6.54 (0.78)	7
Maximilian	1.11 (0.40)	1	Leo	2.34 (1.28)	2	Lotte	6.54 (0.82)	7
Thomas	1.11 (0.40)	1	Robin	2.37 (1.42)	2	Yvonne	6.54 (0.85)	7
Johannes	1.14 (0.36)	1	Milan	2.43 (1.27)	2	Ida	6.57 (0.74)	7
Hugo	1.14 (0.43)	1	Noel	2.74 (1.46)	3	Josephine	6.57 (1.14)	7
Lukas	1.14 (0.43)	1	Gerrit	2.89 (1.76)	3	Amelie	6.60 (1.12)	7
Peter	1.14 (0.43)	1	Ulli	2.91 (1.29)	3	Carolin	6.63 (0.77)	7
Matteo	1.17 (0.45)	1	Lovis	2.94 (1.30)	3	Henriette	6.66 (0.80)	7
Oliver	1.17 (0.45)	1	Florin	3.11 (1.62)	3	Ella	6.66 (0.97)	7
Felix	1.20 (0.47)	1	Toni	3.14 (1.54)	4	Elisabeth	6.66 (1.08)	7
Patrick	1.20 (0.53)	1	Tomke	3.17 (1.54)	4	Marlene	6.69 (0.58)	7
Anton	1.20 (0.58)	1	Renée	3.23 (1.29)	4	Ina	6.69 (0.68)	7
Oskar	1.23 (0.55)	1	Sam	3.31 (1.18)	4	Luisa	6.69 (1.08)	7
Sebastian	1.23 (0.65)	1	Bente	3.37 (1.55)	4	Selina	6.69 (1.08)	7
Erik	1.26 (0.56)	1	Jean	3.43 (1.42)	4	Jasmin	6.71 (0.57)	7
Benedikt	1.26 (0.66)	1	Luca	3.46 (1.60)	4	Greta	6.74 (0.56)	7
Konstantin	1.26 (0.66)	1	Sascha	3.46 (1.70)	4	Lara	6.74 (0.61)	7
Fabian	1.26 (0.70)	1	Mika	3.66 (1.24)	4	Emma	6.74 (0.89)	7
Benjamin	1.26 (0.92)	1	Marlin	3.66 (1.28)	4	Alina	6.77 (0.65)	7
Hans	1.26 (1.04)	1	Jona	3.80 (1.94)	4	Lea	6.77 (1.03)	7
Philipp	1.26 (1.07)	1	Quinn	3.83 (1.60)	4	Maja	6.80 (0.47)	7
Daniel	1.29 (0.62)	1	Charlie	3.97 (1.32)	4	Charlotte	6.80 (0.58)	7
Michael	1.31 (0.68)	1	Marian	4.06 (2.01)	4	Antonia	6.83 (0.38)	7
Timo	1.34 (0.76)	1	Jamie	4.11 (1.02)	4	Marie	6.83 (0.38)	7
Karl	1.34 (1.11)	1	Maxime	4.23 (1.68)	4	Fiona	6.83 (0.45)	7
Adrian	1.37 (0.73)	1	Romy	4.71 (1.60)	4	Hanna	6.83 (0.45)	7
Benno	1.40 (0.69)	1	Kim	4.74 (1.04)	4	Julia	6.83 (0.45)	7
Julian	1.40 (1.17)	1	Sidney	4.74 (1.42)	4	Frieda	6.83 (0.51)	7
Raphael	1.46 (0.89)	1	Elia	4.74 (1.67)	4	Emilia	6.86 (0.36)	7
Florian	1.46 (1.44)	1	Eike	4.80 (1.92)	5	Lina	6.86 (0.36)	7
Finn	1.49 (0.82)	1	Benja	4.91 (1.27)	5	Carla	6.86 (0.43)	7
Hannes	1.51 (0.95)	1	Daniele	4.94 (1.97)	5	Martha	6.86 (0.43)	7
Clemens	1.51 (1.07)	1	Dominique	4.97 (1.92)	5	Lena	6.89 (0.32)	7
Simon	1.51 (1.22)	1	Janne	5.20 (1.55)	5	Leonie	6.89 (0.32)	7
Tim	1.51 (1.46)	1	Kaya	5.31 (1.57)	6	Mia	6.89 (0.32)	7
Jan	1.54 (0.98)	1	Michele	5.63 (1.44)	6	Rosa	6.89 (0.40)	7
Valentin	1.54 (1.17)	1	Juna	5.74 (1.38)	6	Anna	6.91 (0.28)	7
Linus	1.57 (0.88)	1	Andrea	5.91 (1.44)	7	Clara	6.91 (0.28)	7
Emil	1.63 (1.24)	1	Sanja	5.94 (1.35)	6	Mathilda	6.91 (0.28)	7
Kilian	1.66 (0.97)	1	Jule	6.00 (1.37)	7	Sophia	6.91 (0.28)	7
Mats	1.66 (1.03)	1	Alma	6.17 (0.98)	6	Johanna	6.94 (0.24)	7
Damian	1.74 (0.92)	1	Nele	6.17 (1.56)	7	Katharina	6.94 (0.24)	7
Marlon	1.74 (1.09)	1	Mila	6.23 (1.11)	7			

Table A.2: List of filler items (role names translated from English (Kennison & Trofe, 2003)) in ascending mean gender rating order

filler items	mean rating	filler items	mean rating
Kellnerin	1.375	Psychiater	4.050
Stabturnerin	1.400	Schriftsteller	4.150
Balletttänzerin	1.525	Gastwirt	4.250
Flugbegleiterin	1.675	Astrologe	4.350
Stepptänzerin	1.700	Versicherungsvertreter	4.450
Cheerleaderin	1.875	Pharmazeut	4.550
Babysitterin	1.900	Statistiker	4.625
Flugbegleiterin	2.025	Physiker	4.750
Haushälterin	2.075	Professor	4.850
Tanzlehrerin	2.150	Chiropraktiker	4.950
Eiskunstläuferin	2.200	Diplomat	5.050
Stripperin	2.200	Schuldirektor	5.150
Grundschullehrerin	2.250	Zahnarzt	5.275
Bibliothekarin	2.325	Architekt	5.325
Tänzerin	2.450	Politiker	5.450
Turnerin	2.500	Bestattungsunternehmer	5.550
Ernährungsberaterin	2.675	Förster	5.625
Kolumnistin	2.700	Astronaut	5.750
Telefonistin	2.775	Pfandleiher	5.850
Masseurin	2.925	Bauunternehmer	5.925
Bankkassiererin	3.000	Stellvertreter	6.050
Sozialarbeiterin	3.075	Fischer	6.150
Reiseveranstalterin	3.100	Wärter	6.200
Beratungslehrerin	3.225	Schweißer	6.225
Immobilienmaklerin	3.350	Autoverkäufer	6.250
Schulpsychologin	3.450	Barbier	6.325
Kassiererin	3.550	Dachdecker	6.375
Psychologin	3.775	Brunnenbohrer	6.400
Physiotherapeutin	3.875	Wrestler	6.575
Künstlerin	3.925	Kollege	6.700

Appendix B

Model Outputs

B.1 Summary Outputs of lmers with participant_mm_grouping(_nonAmb) as the Main Predictor of Interest

Region 01

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos01_ordNorm ~ participant_mm_grouping_nonAmb + (1 | participant) +
          trial_index_z + list + pro + item_freq_z + participant_gender +      participant_age_z
Data: df_2k
Subset: abs(scale(resid(lmer_2k_P1))) < 2.5

REML criterion at convergence: 33984.7

Scaled residuals:
    Min     1Q   Median     3Q     Max 
-2.9129 -0.5675 -0.1941  0.3190  4.6841

Random effects:
Groups      Name        Variance Std.Dev.
participant (Intercept) 33489     183.0
Residual            39707     199.3
Number of obs: 2525, groups: participant, 61

Fixed effects:
                                         Estimate Std. Error       df t value Pr(>|t|)    
(Intercept)                         621.829    76.868     50.911  8.090 1.07e-10 ***
participant_mm_grouping_nonAmbAmbiguous 11.065     9.800  2463.035  1.129   0.259  
trial_index_z                        -28.677    4.002  2462.280 -7.165 1.02e-12 ***
list2                                -81.750   86.323    50.530 -0.947   0.348  
list3                                -3.812    88.526    50.749 -0.043   0.966  
list4                                -80.982   87.663    50.557 -0.924   0.360  
list5                                29.789    82.910    50.434  0.359   0.721 
```

```

list6                         -43.384    86.020   50.519  -0.504   0.616
proSie                        -3.154     7.966 2460.198  -0.396   0.692
item_freq_z                   -18.752    3.982 2460.112  -4.709 2.63e-06 ***
participant_genderkeineAngabe 45.520    95.658   50.553   0.476   0.636
participant_gendernb          -189.217   147.730  51.234  -1.281   0.206
participant_genderw            12.679    57.528   50.543   0.220   0.826
participant_age_z              -18.121    26.503   50.465  -0.684   0.497
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```

Region 02

```

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos02_ordNorm ~ participant_mm_grouping_nonAmb + (1 | participant) +
           trial_index_z + list + pro + item_freq_z + participant_gender +      participant_age_z
Data: df_2k
Subset: abs(scale(resid(lmer_2k_P2))) < 2.5

```

REML criterion at convergence: 31305.2

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.0681	-0.5775	-0.0970	0.4146	6.4050

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	15191	123.3
Residual		12791	113.1

Number of obs: 2537, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	454.215	51.515	51.443	8.817	7.31e-12 ***
participant_mm_grouping_nonAmbiguous	5.419	5.547	2474.504	0.977	0.329
trial_index_z	-28.859	2.264	2473.906	-12.748	< 2e-16 ***
list2	-24.066	57.885	51.174	-0.416	0.679
list3	-14.372	59.341	51.325	-0.242	0.810
list4	-29.871	58.773	51.165	-0.508	0.613
list5	-2.662	55.604	51.107	-0.048	0.962
list6	2.335	57.677	51.148	0.040	0.968
proSie	-1.274	4.509	2472.756	-0.282	0.778
item_freq_z	-9.949	2.246	2472.555	-4.430	9.82e-06 ***
participant_genderkeineAngabe	65.465	64.164	51.262	1.020	0.312
participant_gendernb	-140.666	98.976	51.710	-1.421	0.161
participant_genderw	1.340	38.583	51.223	0.035	0.972
participant_age_z	3.358	17.772	51.112	0.189	0.851

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Region 04

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos04_ordNorm ~ participant_mm_grouping + (1 | participant) +
         trial_index_z + list + pro + item_freq_z + participant_gender +      participant_age_z
Data: df_2k
Subset: abs(scale(resid(lmer_2k_P4))) < 2.5
```

REML criterion at convergence: 32198.3

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.4120	-0.5828	-0.1573	0.3580	5.8142

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	15335	123.8
Residual		18781	137.0

Number of obs: 2533, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	519.444	52.101	51.534	9.970	1.27e-13 ***
participant_mm_groupingAmbiguous	9.974	7.287	2469.977	1.369	0.171174
participant_mm_groupingMismatch	20.854	6.233	2468.210	3.346	0.000833 ***
trial_index_z	-35.306	2.746	2469.466	-12.857	< 2e-16 ***
list2	10.620	58.441	50.908	0.182	0.856525
list3	-46.845	59.919	51.083	-0.782	0.437938
list4	-53.815	59.335	50.890	-0.907	0.368701
list5	17.410	56.128	50.804	0.310	0.757690
list6	-12.973	58.229	50.873	-0.223	0.824594
proSie	11.919	5.468	2467.561	2.180	0.029359 *
item_freq_z	4.770	2.739	2467.400	1.741	0.081744 .
participant_genderkeineAngabe	50.811	64.774	50.972	0.784	0.436416
participant_gendernb	-210.277	100.057	51.707	-2.102	0.040482 *
participant_genderw	7.628	38.945	50.910	0.196	0.845502
participant_age_z	-7.937	17.941	50.820	-0.442	0.660062

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '

Region 05

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos05_ordNorm ~ participant_mm_grouping + (1 | participant) +
         trial_index_z + list + pro + item_freq_z + participant_gender +      participant_age_z
```

```

Data: df_2k
Subset: abs(scale(resid(lmer_2k_P5))) < 2.5

REML criterion at convergence: 29638

Scaled residuals:
    Min     1Q Median     3Q    Max
-3.2399 -0.5967 -0.1285  0.4068  5.7350

Random effects:
Groups      Name        Variance Std.Dev.
participant (Intercept) 8279      90.99
Residual                 6868      82.87
Number of obs: 2529, groups: participant, 61

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) 396.0941   38.0599 51.6049 10.407 2.82e-14 ***
participant_mm_groupingAmbiguous 9.2099    4.4105 2465.5289  2.088  0.0369 *
participant_mm_groupingMismatch 9.9875    3.7745 2463.9620  2.646  0.0082 **
trial_index_z      -28.8822   1.6648 2464.8351 -17.349 < 2e-16 ***
list2           -9.9963   42.7294 51.1606 -0.234  0.8160
list3           -32.1039   43.7995 51.2909 -0.733  0.4669
list4           -9.9936   43.3851 51.1523 -0.230  0.8187
list5           1.3020   41.0405 51.0675  0.032  0.9748
list6           0.8440   42.5746 51.1274  0.020  0.9843
proSie          7.3318   3.3095 2463.5969  2.215  0.0268 *
item_freq_z      -0.5517   1.6634 2463.4847 -0.332  0.7401
participant_genderkeineAngabe 51.2046   47.3478 51.1758  1.081  0.2846
participant_gendernb      -101.1138   73.0698 51.7193 -1.384  0.1724
participant_genderw       -0.6518   28.4762 51.1749 -0.023  0.9818
participant_age_z         6.2821   13.1185 51.0918  0.479  0.6341
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Region 06

```

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos06_ordNorm ~ participant_mm_grouping + (1 | participant) +
          trial_index_z + list + pro + item_freq_z + participant_gender +      participant_age_z
Data: df_2k
Subset: abs(scale(resid(lmer_2k_P6))) < 2.5

```

REML criterion at convergence: 29837.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

-3.1018 -0.5761 -0.0985 0.4199 7.9400

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	8761	93.60
Residual		7298	85.43

Number of obs: 2533, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	394.3257	39.1551	51.5276	10.071	9.02e-14 ***
participant_mm_groupingAmbiguous	5.8690	4.5594	2469.3948	1.287	0.198
participant_mm_groupingMismatch	4.4996	3.8810	2467.8393	1.159	0.246
trial_index_z	-27.8265	1.7078	2468.8577	-16.294	< 2e-16 ***
list2	-13.1964	43.9581	51.0800	-0.300	0.765
list3	-44.6751	45.0554	51.1934	-0.992	0.326
list4	-16.1470	44.6344	51.0795	-0.362	0.719
list5	-0.3907	42.2222	50.9943	-0.009	0.993
list6	-16.8791	43.7971	51.0383	-0.385	0.702
proSie	2.6734	3.4092	2467.5347	0.784	0.433
item_freq_z	-1.9471	1.7102	2467.3971	-1.139	0.255
participant_genderkeineAngabe	71.3282	48.6987	51.0502	1.465	0.149
participant_gendernb	-101.1273	75.1632	51.6160	-1.345	0.184
participant_genderw	0.5864	29.2931	51.0802	0.020	0.984
participant_age_z	4.1795	13.4981	51.0471	0.310	0.758

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '
					1

Region 07

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos07_ordNorm ~ participant_mm_grouping + (1 | participant) +
trial_index_z + list + pro + item_freq_z + participant_gender + participant_age_z
Data: df_2k
Subset: abs(scale(resid(lmer_2k_P7))) < 2.5

REML criterion at convergence: 31253.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.2401	-0.5625	-0.1141	0.4115	6.0661

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	13323	115.4
Residual		12161	110.3

Number of obs: 2544, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	437.5785	48.3325	51.6171	9.054	3.07e-12 ***
participant_mm_groupingAmbiguous	8.8789	5.8538	2480.4623	1.517	0.129
participant_mm_groupingMismatch	-6.1003	5.0077	2479.0712	-1.218	0.223
trial_index_z	-29.2381	2.2045	2480.0923	-13.263	< 2e-16 ***
list2	-5.0417	54.2528	51.1370	-0.093	0.926
list3	-33.9801	55.6161	51.2834	-0.611	0.544
list4	-15.9497	55.0904	51.1474	-0.290	0.773
list5	-1.7805	52.1143	51.0670	-0.034	0.973
list6	10.5570	54.0605	51.1197	0.195	0.846
proSie	0.4232	4.3911	2478.6049	0.096	0.923
item_freq_z	-3.5685	2.2022	2478.5525	-1.620	0.105
participant_genderkeineAngabe	79.6205	60.1143	51.1433	1.324	0.191
participant_gendernb	-126.1331	92.7888	51.7238	-1.359	0.180
participant_genderw	3.5363	36.1559	51.1518	0.098	0.922
participant_age_z	10.4479	16.6577	51.0851	0.627	0.533

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

B.2 Summary Outputs of lmers with participant_itemPro_mm_num as the Main Predictor of Interest

Region 01

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos01_ordNorm ~ participant_itemPro_mm_num + (1 | participant) +
trial_index_z + list + pro + item_freq_z + participant_gender + participant_age_z
Data: df_2k
Subset: abs(scale(resid(lmer_2k_7step_P1_num))) < 2.5

REML criterion at convergence: 33958.9

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.8988	-0.5755	-0.1922	0.3247	4.7351

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	33564	183.2
Residual		39664	199.2

Number of obs: 2523, groups: participant, 61

Fixed effects:

Estimate	Std. Error	df	t value	Pr(> t)
----------	------------	----	---------	----------

(Intercept)	616.966	77.034	51.133	8.009	1.38e-10	***
participant_itemPro_mm_num	2.706	1.577	2458.581	1.716	0.0864	.
trial_index_z	-29.098	3.998	2460.280	-7.279	4.52e-13	***
list2	-81.801	86.415	50.529	-0.947	0.3483	
list3	-2.490	88.617	50.742	-0.028	0.9777	
list4	-80.679	87.756	50.555	-0.919	0.3623	
list5	30.609	82.998	50.433	0.369	0.7138	
list6	-43.990	86.112	50.519	-0.511	0.6117	
proSie	-3.946	7.965	2458.172	-0.495	0.6203	
item_freq_z	-19.662	3.993	2458.085	-4.924	9.05e-07	***
participant_genderkeineAngabe	45.607	95.759	50.552	0.476	0.6359	
participant_gendernb	-187.928	147.886	51.233	-1.271	0.2096	
participant_genderw	12.691	57.590	50.544	0.220	0.8265	
participant_age_z	-18.094	26.531	50.464	-0.682	0.4984	

Signif. codes:	0 *** 0.001 ** 0.01 * 0.05 . 0.1 ' 1					

Region 02

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
 Formula: rt_pos01_ordNorm ~ participant_itemPro_mm_num + (1 | participant) +
 trial_index_z + list + pro + item_freq_z + participant_gender + participant_age_z
 Data: df_2k
 Subset: abs(scale(resid(lmer_2k_7step_P1_num))) < 2.5

REML criterion at convergence: 33958.9

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.8988	-0.5755	-0.1922	0.3247	4.7351

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	33564	183.2
Residual		39664	199.2

Number of obs: 2523, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	616.966	77.034	51.133	8.009	1.38e-10 ***
participant_itemPro_mm_num	2.706	1.577	2458.581	1.716	0.0864 .
trial_index_z	-29.098	3.998	2460.280	-7.279	4.52e-13 ***
list2	-81.801	86.415	50.529	-0.947	0.3483
list3	-2.490	88.617	50.742	-0.028	0.9777
list4	-80.679	87.756	50.555	-0.919	0.3623
list5	30.609	82.998	50.433	0.369	0.7138
list6	-43.990	86.112	50.519	-0.511	0.6117

proSie	-3.946	7.965	2458.172	-0.495	0.6203
item_freq_z	-19.662	3.993	2458.085	-4.924	9.05e-07 ***
participant_genderkeineAngabe	45.607	95.759	50.552	0.476	0.6359
participant_gendernb	-187.928	147.886	51.233	-1.271	0.2096
participant_genderw	12.691	57.590	50.544	0.220	0.8265
participant_age_z	-18.094	26.531	50.464	-0.682	0.4984
<hr/>					

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Region 04

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
 Formula: rt_pos04_ordNorm ~ participant_itemPro_mm_num + (1 | participant) +
 trial_index_z + list + pro + item_freq_z + participant_gender + participant_age_z
 Data: df_2k
 Subset: abs(scale(resid(lmer_2k_7step_P4_num))) < 2.5

REML criterion at convergence: 32207.3

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.4095	-0.5803	-0.1559	0.3624	5.8187

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	15330	123.8
Residual		18773	137.0

Number of obs: 2533, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	518.928	52.088	51.514	9.963	1.31e-13 ***
participant_itemPro_mm_num	3.647	1.082	2469.091	3.370	0.000764 ***
trial_index_z	-35.299	2.744	2470.461	-12.865	< 2e-16 ***
list2	10.648	58.430	50.904	0.182	0.856125
list3	-46.920	59.907	51.074	-0.783	0.437118
list4	-53.749	59.325	50.888	-0.906	0.369202
list5	17.375	56.118	50.800	0.310	0.758118
list6	-12.995	58.219	50.870	-0.223	0.824263
proSie	11.942	5.466	2468.557	2.185	0.029000 *
item_freq_z	4.789	2.735	2468.424	1.751	0.080057 .
participant_genderkeineAngabe	50.728	64.763	50.972	0.783	0.437081
participant_gendernb	-210.165	100.040	51.707	-2.101	0.040552 *
participant_genderw	7.608	38.938	50.910	0.195	0.845860
participant_age_z	-7.950	17.937	50.819	-0.443	0.659501
<hr/>					

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Region 05

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos05_ordNorm ~ participant_itemPro_mm_num + (1 | participant) +
          trial_index_z + list + pro + item_freq_z + participant_gender +      participant_age_z
Data: df_2k
Subset: abs(scale(resid(lmer_2k_7step_P5_num))) < 2.5
```

REML criterion at convergence: 29635.1

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.2390	-0.5928	-0.1276	0.4070	5.7263

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	8257	90.87
Residual		6868	82.87

Number of obs: 2528, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	396.5307	38.0068	51.5830	10.433	2.6e-14 ***
participant_itemPro_mm_num	1.7919	0.6554	2463.9168	2.734	0.0063 **
trial_index_z	-28.9443	1.6634	2464.8366	-17.400	< 2e-16 ***
list2	-9.7705	42.6725	51.1516	-0.229	0.8198
list3	-31.7543	43.7404	51.2783	-0.726	0.4712
list4	-9.9428	43.3276	51.1446	-0.229	0.8194
list5	1.5177	40.9856	51.0574	0.037	0.9706
list6	1.0704	42.5177	51.1173	0.025	0.9800
proSie	7.2011	3.3095	2463.5947	2.176	0.0297 *
item_freq_z	-0.6715	1.6623	2463.4982	-0.404	0.6863
participant_genderkeineAngabe	51.2434	47.2852	51.1687	1.084	0.2836
participant_gendernb	-101.0151	72.9737	51.7137	-1.384	0.1722
participant_genderw	-0.6304	28.4386	51.1680	-0.022	0.9824
participant_age_z	6.2159	13.1011	51.0842	0.474	0.6372

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Region 06

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos06_ordNorm ~ participant_itemPro_mm_num + (1 | participant) +
          trial_index_z + list + pro + item_freq_z + participant_gender +      participant_age_z
Data: df_2k
Subset: abs(scale(resid(lmer_2k_7step_P6_num))) < 2.5
```

REML criterion at convergence: 29834.8

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.0695	-0.5768	-0.0987	0.4184	7.9219

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	8759	93.59
Residual		7298	85.43

Number of obs: 2532, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	394.2295	39.1493	51.5150	10.070	9.08e-14 ***
participant_itemPro_mm_num	1.0201	0.6738	2467.7882	1.514	0.130
trial_index_z	-27.8901	1.7070	2468.8244	-16.338	< 2e-16 ***
list2	-13.0445	43.9540	51.0787	-0.297	0.768
list3	-44.4080	45.0504	51.1886	-0.986	0.329
list4	-15.9631	44.6302	51.0782	-0.358	0.722
list5	-0.2250	42.2179	50.9917	-0.005	0.996
list6	-16.7049	43.7927	51.0357	-0.381	0.704
proSie	2.6372	3.4097	2467.5402	0.773	0.439
item_freq_z	-2.0662	1.7080	2467.4192	-1.210	0.227
participant_genderkeineAngabe	71.3530	48.6947	51.0511	1.465	0.149
participant_gendernb	-100.9516	75.1567	51.6165	-1.343	0.185
participant_genderw	0.5789	29.2907	51.0818	0.020	0.984
participant_age_z	4.1621	13.4970	51.0473	0.308	0.759

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '
	1				

Region 07

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: rt_pos07_ordNorm ~ participant_itemPro_mm_num + (1 | participant) + trial_index_z + list + pro + item_freq_z + participant_gender + participant_age_z
Data: df_2k
Subset: abs(scale(resid(lmer_2k_7step_P7_num))) < 2.5

REML criterion at convergence: 31317.7

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.1498	-0.5673	-0.1123	0.4103	6.1519

Random effects:

```

Groups      Name      Variance Std.Dev.
participant (Intercept) 13208     114.9
Residual             12189     110.4
Number of obs: 2548, groups: participant, 61

Fixed effects:
                                         Estimate Std. Error      df t value Pr(>|t|)
(Intercept)                      438.9983   48.1278    51.6209   9.122 2.41e-12 ***
participant_itemPro_mm_num      -0.7904    0.8693 2483.9781  -0.909  0.3633
trial_index_z                   -29.4667   2.2056 2485.1244 -13.360 < 2e-16 ***
list2                           -4.4780   54.0252   51.1487 -0.083  0.9343
list3                           -32.8007   55.3816   51.2909 -0.592  0.5563
list4                           -16.0708   54.8587   51.1572 -0.293  0.7707
list5                           -1.1788   51.8958   51.0793 -0.023  0.9820
list6                           11.1296   53.8338   51.1317  0.207  0.8370
proSie                          0.0731   4.3921 2483.6118  0.017  0.9867
item_freq_z                     -3.7904   2.1993 2483.5691 -1.723  0.0849 .
participant_genderkeineAngabe  79.9916   59.8630   51.1583  1.336  0.1874
participant_gendernb            -125.8996  92.4037   51.7452 -1.362  0.1789
participant_genderw              3.6937   36.0046   51.1658  0.103  0.9187
participant_age_z                10.2768  16.5878   51.0962  0.620  0.5383
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

B.3 Summary Outputs of the *sie*-subset Analysis

Region 04

```

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos04_ordNorm ~ participant_mm_grouping + (1 | participant) +
          trial_index_z + list + participant_gender + item_freq_z +      participant_age_z
Data: df_2kSie
Subset: abs(scale(resid(lmer_2kSie_P4))) < 2.5

```

REML criterion at convergence: 16267.5

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.2570	-0.5845	-0.1587	0.3441	5.8896

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	14894	122.0
Residual		19687	140.3

Number of obs: 1275, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	538.472	52.298	51.785	10.296	3.95e-14 ***
participant_mm_groupingMismatch	18.084	9.006	1211.842	2.008	0.0449 *
participant_mm_groupingnAmbiguous	12.553	10.698	1215.132	1.173	0.2409
trial_index_z	-33.091	3.833	1214.924	-8.633	< 2e-16 ***
list2	1.795	58.659	51.154	0.031	0.9757
list3	-46.744	60.148	51.324	-0.777	0.4406
list4	-64.672	59.452	50.782	-1.088	0.2818
list5	19.046	56.290	50.872	0.338	0.7365
list6	-12.479	58.378	50.876	-0.214	0.8316
participant_genderkeineAngabe	17.741	65.054	51.325	0.273	0.7862
participant_gendernb	-214.537	100.711	52.537	-2.130	0.0379 *
participant_genderw	5.702	39.024	50.810	0.146	0.8844
item_freq_z	3.677	3.858	1210.991	0.953	0.3408
participant_age_z	-10.083	17.974	50.694	-0.561	0.5773

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Region 05

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos05_ordNorm ~ participant_mm_grouping + (1 | participant) +
trial_index_z + list + participant_gender + item_freq_z + participant_age_z
Data: df_2kSie
Subset: abs(scale(resid(lmer_2kSie_P5))) < 2.5

REML criterion at convergence: 14958.5

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.8076	-0.5935	-0.1364	0.4328	4.6709

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	8281	91.00
Residual		6914	83.15

Number of obs: 1274, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	412.0864	38.4866	51.9035	10.707	9.53e-15 ***
participant_mm_groupingMismatch	5.6379	5.3309	1210.4709	1.058	0.2905
participant_mm_groupingnAmbiguous	10.7827	6.3644	1212.8970	1.694	0.0905 .
trial_index_z	-28.5409	2.2771	1212.4964	-12.534	< 2e-16 ***
list2	-14.1410	43.2115	51.4750	-0.327	0.7448
list3	-41.7093	44.3080	51.6621	-0.941	0.3509

list4	-14.3424	43.8239	51.2315	-0.327	0.7448
list5	-5.4041	41.4726	51.2289	-0.130	0.8968
list6	-5.7447	43.0239	51.2916	-0.134	0.8943
participant_genderkeineAngabe	45.3702	47.8595	51.3936	0.948	0.3476
participant_gendernb	-104.8852	74.0566	52.4969	-1.416	0.1626
participant_genderw	-3.8929	28.7682	51.2794	-0.135	0.8929
item_freq_z	-0.9029	2.2815	1209.8660	-0.396	0.6924
participant_age_z	6.4621	13.2486	51.1311	0.488	0.6278

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Region 06

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
 Formula: rt_pos06_ordNorm ~ participant_mm_grouping + (1 | participant) +
 trial_index_z + list + participant_gender + item_freq_z + participant_age_z
 Data: df_2kSie
 Subset: abs(scale(resid(lmer_2kSie_P6))) < 2.5

REML criterion at convergence: 14882.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.4480	-0.6057	-0.0869	0.4710	4.9591

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	8616	92.82
Residual		6360	79.75

Number of obs: 1276, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	399.6776	39.1599	51.5804	10.206	5.61e-14 ***
participant_mm_groupingMismatch	-0.4683	5.1224	1212.0680	-0.091	0.927
participant_mm_groupingnAmbiguous	8.8631	6.0661	1214.0586	1.461	0.144
trial_index_z	-28.8591	2.1684	1213.9126	-13.309	< 2e-16 ***
list2	-17.3054	43.9723	51.1767	-0.394	0.696
list3	-51.4845	45.0774	51.3150	-1.142	0.259
list4	-15.0105	44.6046	50.9747	-0.337	0.738
list5	-14.6770	42.2123	50.9771	-0.348	0.730
list6	-13.8612	43.7787	50.9812	-0.317	0.753
participant_genderkeineAngabe	51.0406	48.6828	51.0135	1.048	0.299
participant_gendernb	-93.1844	75.2525	51.8903	-1.238	0.221
participant_genderw	4.6459	29.2776	51.0016	0.159	0.875
item_freq_z	0.8493	2.1888	1211.6367	0.388	0.698
participant_age_z	3.3596	13.4875	50.9182	0.249	0.804

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Region 07

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos07_ordNorm ~ participant_mm_grouping + (1 | participant) +
trial_index_z + list + participant_gender + item_freq_z + participant_age_z
Data: df_2kSie
Subset: abs(scale(resid(lmer_2kSie_P7))) < 2.5

REML criterion at convergence: 15576.5

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.6852	-0.6047	-0.1120	0.4332	4.6484

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	11812	108.7
Residual		11125	105.5

Number of obs: 1276, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	447.324	46.091	51.886	9.705	2.96e-13 ***
participant_mm_groupingMismatch	-11.435	6.757	1212.731	-1.692	0.0909 .
participant_mm_groupingnAmbiguous	4.559	8.057	1214.785	0.566	0.5716
trial_index_z	-28.993	2.878	1214.793	-10.076	< 2e-16 ***
list2	-10.537	51.735	51.403	-0.204	0.8394
list3	-40.843	53.053	51.603	-0.770	0.4449
list4	-20.677	52.481	51.206	-0.394	0.6952
list5	-10.871	49.664	51.198	-0.219	0.8276
list6	4.401	51.505	51.196	0.085	0.9322
participant_genderkeineAngabe	57.598	57.295	51.301	1.005	0.3195
participant_gendernb	-127.220	88.642	52.362	-1.435	0.1572
participant_genderw	1.582	34.441	51.196	0.046	0.9635
item_freq_z	-3.139	2.898	1211.964	-1.083	0.2789
participant_age_z	10.604	15.865	51.102	0.668	0.5069

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation matrix not shown by default, as p = 14 > 12.

Use print(x, correlation=TRUE) or
vcov(x) if you need it

B.4 Summary Outputs of the *er*-subset Analysis

Region 04

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos04_ordNorm ~ participant_mm_grouping + (1 | participant) +
          trial_index_z + list + participant_gender + item_freq_z +      participant_age_z
Data: df_2kEr
Subset: abs(scale(resid(lmer_2kEr_P4))) < 2.5
```

REML criterion at convergence: 15995.2

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.6150	-0.5744	-0.1513	0.3632	5.6427

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	14343	119.8
Residual		18436	135.8

Number of obs: 1260, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	516.481	51.248	51.511	10.078	8.84e-14 ***
participant_mm_groupingMismatch	18.886	8.860	1197.459	2.132	0.0332 *
participant_mm_groupingnAmbiguous	7.669	10.162	1198.780	0.755	0.4506
trial_index_z	-38.386	4.114	1196.702	-9.330	< 2e-16 ***
list2	17.818	57.397	50.578	0.310	0.7575
list3	-43.428	58.942	51.071	-0.737	0.4646
list4	-45.410	58.387	50.946	-0.778	0.4403
list5	7.361	55.117	50.447	0.134	0.8943
list6	-12.177	57.259	50.796	-0.213	0.8324
participant_genderkeineAngabe	66.466	63.669	50.807	1.044	0.3015
participant_gendernb	-206.233	98.796	52.460	-2.087	0.0417 *
participant_genderw	10.788	38.320	50.952	0.282	0.7794
item_freq_z	5.518	4.002	1196.295	1.379	0.1682
participant_age_z	-5.418	17.642	50.743	-0.307	0.7600

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '

Region 05

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos05_ordNorm ~ participant_mm_grouping + (1 | participant) +
          trial_index_z + list + participant_gender + item_freq_z +      participant_age_z
Data: df_2kEr
```

```
Subset: abs(scale(resid(lmer_2kEr_P5))) < 2.5
```

```
REML criterion at convergence: 14712.1
```

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.8294	-0.5819	-0.1349	0.3852	5.7557

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	8339	91.32
Residual		6844	82.73

Number of obs: 1254, groups: participant, 61

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	389.116	38.593	51.705	10.083	8.34e-14 ***
participant_mm_groupingMismatch	12.911	5.417	1191.061	2.383	0.0173 *
participant_mm_groupingnAmbiguous	5.145	6.210	1191.922	0.829	0.4075
trial_index_z	-31.105	2.505	1190.332	-12.415	< 2e-16 ***
list2	-10.503	43.294	51.103	-0.243	0.8093
list3	-23.704	44.421	51.424	-0.534	0.5959
list4	-8.617	44.013	51.343	-0.196	0.8455
list5	5.999	41.569	50.939	0.144	0.8858
list6	1.804	43.160	51.177	0.042	0.9668
participant_genderkeineAngabe	56.450	48.013	51.284	1.176	0.2451
participant_gendernb	-94.421	74.249	52.244	-1.272	0.2091
participant_genderw	6.081	28.887	51.360	0.211	0.8341
item_freq_z	2.526	2.439	1190.373	1.036	0.3005
participant_age_z	6.077	13.303	51.203	0.457	0.6498

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Region 06

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos06_ordNorm ~ participant_mm_grouping + (1 | participant) +
   trial_index_z + list + participant_gender + item_freq_z +      participant_age_z
Data: df_2kEr
Subset: abs(scale(resid(lmer_2kEr_P6))) < 2.5
```

REML criterion at convergence: 14904.1

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.0220	-0.5425	-0.0946	0.3757	7.5783

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	8478	92.07
Residual		7878	88.76
Number of obs: 1256, groups: participant, 61			

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	389.806	39.028	51.699	9.988	1.16e-13 ***
participant_mm_groupingMismatch	8.015	5.788	1193.190	1.385	0.1664
participant_mm_groupingnAmbiguous	5.218	6.679	1194.029	0.781	0.4348
trial_index_z	-27.154	2.695	1192.648	-10.075	< 2e-16 ***
list2	-11.606	43.769	51.032	-0.265	0.7920
list3	-38.639	44.905	51.344	-0.860	0.3935
list4	-23.345	44.488	51.241	-0.525	0.6020
list5	6.320	42.028	50.891	0.150	0.8811
list6	-25.219	43.626	51.073	-0.578	0.5657
participant_genderkeineAngabe	97.129	48.512	51.099	2.002	0.0506 .
participant_gendernb	-106.031	75.153	52.425	-1.411	0.1642
participant_genderw	3.680	29.197	51.241	0.126	0.9002
item_freq_z	-3.356	2.624	1192.424	-1.279	0.2011
participant_age_z	4.340	13.453	51.193	0.323	0.7483

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Region 07

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: rt_pos07_ordNorm ~ participant_mm_grouping + (1 | participant) +
trial_index_z + list + participant_gender + item_freq_z + participant_age_z
Data: df_2kEr
Subset: abs(scale(resid(lmer_2kEr_P7))) < 2.5

REML criterion at convergence: 15691.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.6795	-0.5239	-0.1190	0.3981	5.5802

Random effects:

Groups	Name	Variance	Std.Dev.
participant	(Intercept)	13979	118.2
Residual		13460	116.0
Number of obs: 1266, groups: participant, 61			

Fixed effects:

Estimate	Std. Error	df	t value	Pr(> t)

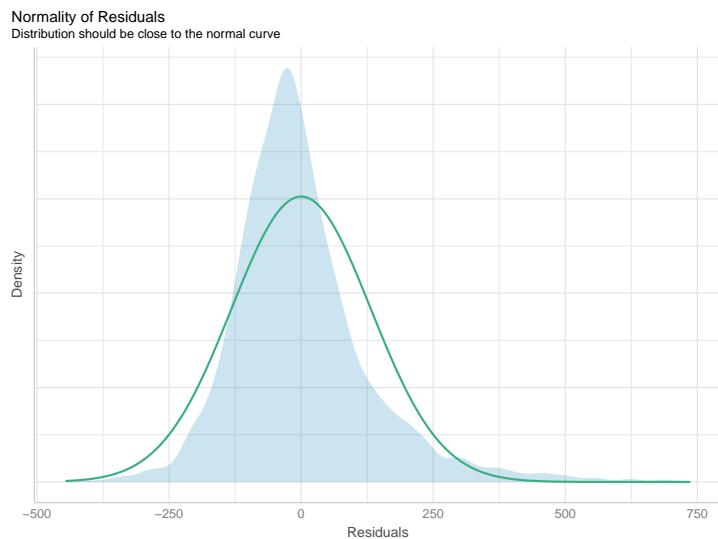
(Intercept)	429.7068	50.1522	51.8585	8.568	1.66e-11	***
participant_mm_groupingMismatch	-2.0577	7.5419	1203.4261	-0.273	0.785	
participant_mm_groupingnAmbiguous	9.5756	8.6979	1204.4663	1.101	0.271	
trial_index_z	-29.8837	3.5155	1202.9331	-8.501	< 2e-16	***
list2	0.5862	56.2229	51.1130	0.010	0.992	
list3	-22.1572	57.7247	51.5749	-0.384	0.703	
list4	-15.5026	57.1630	51.3804	-0.271	0.787	
list5	4.2118	54.0115	51.0636	0.078	0.938	
list6	11.3192	56.0633	51.2424	0.202	0.841	
participant_genderkeineAngabe	99.1805	62.3443	51.2753	1.591	0.118	
participant_gendernb	-126.7349	96.5395	52.5073	-1.313	0.195	
participant_genderw	7.3907	37.5202	51.4078	0.197	0.845	
item_freq_z	-3.5889	3.3980	1202.4648	-1.056	0.291	
participant_age_z	8.2555	17.2745	51.2034	0.478	0.635	

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .	0.1 ' '	1

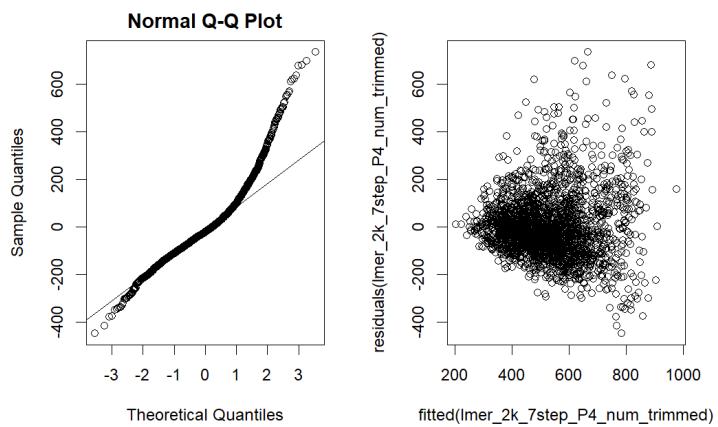
B.5 Model Diagnostics for the Continuous Analysis

Figure B.1: Model diagnostics in the Continuous Analysis

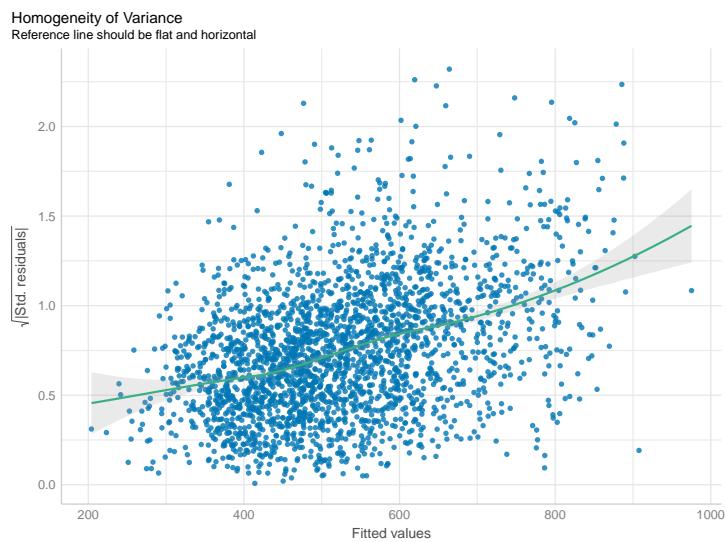
(a) Normal probability plot for the final model on region 04



(b) Q-Q plot and residual plot for the final model on region 04



(c) Linearity and Homoscedasticity plot for the final model on region 04



Appendix C

Glossary

M = grammatical gender masculine

F = grammatical gender feminine

N = grammatical gender neuter

<MALE> = stereotypical gender male

<FEMALE> = stereotypical gender female

<MALE/FEMALE> = stereotypical gender male and female

pos = presentation region (outdated: “position”)