It's Personal and Disgusting: Extra-Linguistic Information in Language Comprehension

by

Isabell Hubert Lyall

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics
University of Alberta

Examining committee:

Juhani Järvikivi, Supervisor
Antti Arppe, Supervisory Committee
Benjamin V. Tucker, Supervisory Committee
Elena Nicoladis, Examiner
Julie Boland, External Examiner

Abstract

This dissertation examines the influence of various extra-linguistic aspects on language comprehension. While language comprehension is generally understood to be influenced by real-world context, and by certain individual difference variables such as the listener's mood, it is unclear how an individual's personality and political views interact with variables inferred about the speaker when understanding language. This dissertation thus investigated how aspects of a listener's identity (namely their personality, political views, and Disgust Sensitivity), combined with aspects inferred about the speaker's identity (specifically, their gender inferred from their voice), influences language comprehension. Additionally, this dissertation presents the first investigation of Disgust Sensitivity within the context of linguistic processing. Disgust Sensitivity is assumed to be a marker of Behavioural Immune System activity, which attempts to protect an organism from pathogens and is thus assumed to correlate with a person's outgroup stigmatization tendencies. To assess the effects of the variables mentioned on both conscious and sub-conscious language comprehension, a multi-methodological array of four psycholinguistic experiments was conducted, using the item rating, self-paced listening, and pupillometry paradigms. Crucially, a portion of the auditory stimuli in each experiment contained one of three types of clashes: Morpho-syntactic errors (such as "He often *walk* his dog..."), semantic anomalies (such as "Dogs often chase *teas*..."), and socio-cultural clashes based on established gender stereotypes (such as "I buy my *bras*..." spoken by a male speaker), and it was especially the listener's responses to these clashes that were of interest. Specifically, the four experiments in this dissertation investigated whether the variables in question modulated language comprehension, and whether the processing of the three clash types was influenced by different variables. It was additionally hypothesized that

Disgust Sensitivity would specifically modulate responses to socio-cultural clashes. Results from the four experiments indicate that personality traits, political values, and Disgust Sensitivity indeed affect language comprehension, but that no one variable affects it across the board. Results are in line with a view of language comprehension that includes anticipation based on contextual factors, and that assigns importance to extra-linguistic variables. Results further suggest that "intra-linguistic information" is not considered separately, in a first step, with the utterance being integrated with extra-linguistic information at a later point; rather, results are compatible with a (constraint-based) one-step model of language comprehension, where all available information is used in anticipation, and in one single step of comprehension. Results are thus broadly supportive of a cognitive linguistic view, and are not at odds with experiential accounts of linguistic representation.

Preface

This thesis is an original work by Isabell Hubert Lyall. The research project which this thesis is a part of has received research ethics approval from the University of Alberta Research Ethics Board 2, project name *Extra-Linguistic Information in Language Comprehension*, `Pro00065357`, 27 June 2016, and *ExtraLing: Pupillometry & Disgust*, `Pro00077213`, 23 January 2018.

The research conducted in this dissertation was completed in collaboration with Dr. Juhani Järvikivi. Dr. Järvikivi assisted in concept formulation, experiment design, and data analysis. Dr. Järvikivi, Dr. Antti Arppe, and Dr. Benjamin V. Tucker assisted in the editing of the manuscript. I was responsible for experiment design, data collection, analysis, and manuscript composition. Undergraduate assistants Kimberly Coleman, Jessica Kondor, Sarah Noga, and Liana Oh assisted with data collection and experiment testing.

Recruitment for Experiment IV (Chapter 5) was supported through a Social Sciences and Humanities Research Council (SSHRC) Partnership Grant, “Words in the World,” 895-2016-1008, to Dr. Järvikivi.

Parts of the research presented in Chapter 5 were presented at the *Annual Meeting of the Cognitive Science Society* (Montréal, QC, 25-27 July 2019), with subsequent publication in the conference proceedings as: Hubert, I., and Järvikivi, J. (2019). *Dark Forces in Language Comprehension: The Case of Neuroticism and Disgust in a Pupillometry Study.* In A.K. Goel, C.M. Seifert, & C. Freksa (Eds.), Proceedings of the 41st Annual Conference of the Cognitive Science Society. I was responsible for data collection and analysis as well as manuscript composition. Dr. Järvikivi was the supervisory author and assisted with concept formation, data analysis, and contributed to manuscript composition.

Further, parts of this dissertation research were presented at conferences, as indicated in footnotes at the beginning of the respective chapters.

Acknowledgements

As is always the case for a work of this size and scope, there are far too many people to acknowledge and give thanks to. I'll try anyways.

Thank you to my doctoral supervisor, Juhani Järvikivi, for his support through the years, which sported a successful blend of guidance and free roam. I cannot possibly find the words to describe how nurturing and encouraging your supervision has been. I could not have wished for a better supervisor.

Thank you also to the members of my supervisory committee, Antti Arppe and Benjamin V. Tucker, for valuable feedback on drafts of both the Thesis Prospectus and this dissertation. Thank you also to Herbert Colston for valuable psychological insights and feedback during Candidacy. Thank-you also to Adriana Hanulíková for extremely valuable feedback on my thesis prospectus and during Candidacy, and to Elena Nicoladis and Julie Boland for highly valuable comments on this manuscript and during the Thesis Defense.

Thank you to my undergraduate research assistants, Kimberly Coleman, Jessica Kondor, Sarah Noga, and Liana Oh, for data collection and experiment testing. Your contributions helped speed up data collection greatly.

I am thankful for the support I have received through the *Words in the World* project, which has enabled broader recruitment for the final experiment in this dissertation, and travel to present my work at various conferences in Canada and abroad.

Many thanks also to Kaidi Lõo for her expert advice on GAMMs and pupillometry, and for many academic and non-academic coffee breaks; to Jacolien van Rij, Yoichi Mukai, Vince Porretta, and Brian Rusk for their comments on all things eye-tracking, pupillometry, and the statistical analysis of the resulting data; to Kaleigh Park, Lauren Rudat, and Hannah Sysak for efficient, supportive, and cheery coordination of lab matters at the CCP; to the members of the CCP for feedback and advice over the years, and to my colleagues at the Department off Linguistics.

Outside of the academic realm, I would like to thank my parents and grandparents; in addition to many other environmental variables, it was their life choices that made it possible for me to go down a path that led to three university degrees. Grandpa, I hope you're watching – you wouldn't believe this!!

Thank you also to Nik and Jay, and Tlats'ux̂, Eskaaya, and Sammy for many, many board game nights and dinners, and to Margaret for our regular weekend brunch meets. They all provided much needed grounding and diversion. You are good people and I am so glad to have you in my life.

Thank you to the board members of the Edmonton Transit Service Advisory Board, Paths for People, and all my urbanist friends and acquaintances. I loved having one of my legs firmly planted in urbanism during this gigantic project, and I have had the fortune of meeting so many fantastic, engaged, and thoughtful people because of it.

Thank you to IPAC Edmonton, the Department of Communications & Engagement at the City of Edmonton, and the Public Service Commission at the Government of Alberta – my two days a week away from academia in 2018 were immensely helpful in broadening my horizons and in meeting lots of fantastic people.

I need to thank Catticus, Scout, and Boo for not deleting any part of this dissertation while walking over my keyboard repeatedly, and with admirable consistence. I'd thank them for their furry (sometimes scratchy) support, too, but they don't seem to care all that much as my written thanks don't come with treats.

Lastly, a world of thanks, and more, to James – I cannot, and do not wish to, imagine what the last few years would have been like without you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The comprehension of linguistic utterances does not only involve the linguistic features, such as words and phrases, that the listener hears. Rather, certain extra-linguistic factors, such as the listener's visual context, world knowledge, or social factors, influence language comprehension processes (Hagoort et al. 2004; Kamide et al. 2003; Sedivy et al. 1999; Traxler 2014; Van Berkum et al. 2005). Specifically, listeners appear to use their experiences in their surroundings, and the knowledge that certain things commonly co-occur with other things, in calculating what they anticipate, or expect, an upcoming segment to be.

As such, recent findings support a model of language comprehension that does not operate in a two-step, syntax-first manner, as proposed by the *standard two-step model of language interpretation* (see e.g. Chomsky 1957; Cutler and Clifton 1999; Grice 1975; Lattner and Friederici 2003); but rather a model that considers prior context and other factors in the same step as syntactic constraints (see e.g. Federmeier 2007; Hagoort et al. 2004; Hagoort and Van Berkum 2007; Levy 2008; Traxler 2014; Van Berkum et al. 2005, 2008; Van Petten and Luka 2012). Prior research, which will be discussed in detail in the following sections (specifically see Section 1.2.2), suggests that world knowledge, and other extra-linguistic factors, seem to influence language comprehension rapidly, that is, at the same time as syntactic constraints.

However, the effects of listener-related and speaker-related variables on language comprehension, in addition to the interplay between the two, are unclear. Hence, this dissertation examines the effects of different extra-linguistic variables on language comprehension. Specifically, four experiments using three different experimental paradigms investigate whether different variables related to the listener's inner state (such as their personality, their political values, and their Disgust Sensitivity), and variables related to the speaker (such as their gender inferred from their voice), are considered in the comprehension of spoken sentences with three different types of clashes: morpho-syntactic errors (such as "He often *walk* his dog..."), semantic anomalies (such as "Dogs often chase *teas*..."), and socio-cultural clashes (such as "I buy my *bras*..." spoken by a male speaker).

As will be shown below, the processing of all three clash types was found to be modulated by extra-linguistic variables; however, different variables seem to affect the processing of different kinds of clashes, and in different interactions with other variables.

In the upcoming sections, we will first discuss different types of extra-linguistic factors that have been found to affect language comprehension, and secondly, how these factors tie in with anticipation in both general cognition and language comprehension.

1.1 Extra-Linguistic Influences on Language Comprehension

It is now generally agreed upon that language comprehension does not operate on a two-step, syntax-first basis, but rather with contextual information being integrated incrementally and rapidly (Kamide et al. 2003; Sedivy et al. 1999; Tanenhaus et al. 1995; Traxler 2014; Van Berkum et al. 2005). Based on experimental results indicating that utterances such as "the girl comforted the clock" can be non-anomalous if the context warrants such an interpretation, Nieuwland and Van Berkum (2006) conclude that "discourse context can immediately overrule local lexical–semantic violations" (Nieuwland and Van Berkum 2006, p. 1098), that there really is no context-free meaning. In a classic two-step model, such an utterance would always have to be anomalous in the first step, as integration with preceding context only happens at a later stage.[1]

"Context" does not necessarily have to be textual; listeners also seem to take the affordances of objects in their surroundings, and their own behavioural goals, into account when searching for a suitable referent (Chambers et al. 2004; Spivey et al. 2002; Tanenhaus et al. 1995). For example, in Tanenhaus et al. (1995)'s visual world eye-tracking experiment, the prepositional phrase "on the towel" that follows "put the apple..." can be interpreted as a modifier (*the apple that is on the towel* as opposed to *the apple that is on the table*, for example), or as the destination (*transfer the apple onto the towel* as opposed to *transfer the*

[1] Please also refer to Section 1.2.2 in this context, for a discussion of linguistic anticipation.

apple into the box, for example). Fixation pattern results showed that, when there was only one apple present in the visual context, “on the towel” was interpreted as the goal; whereas if there were two apples visible, one of which was sitting on a towel, and one of which was not, “on the towel” was interpreted as a modifier. Similar findings emerged from Chambers et al. (2004): Participants were given instructions to move items around, for example into bowls, as their gaze was recorded. In a sentence like “put the egg in the bowl over the flour,” “in the bowl” can either be interpreted as a modifier (*the egg in the bowl* as opposed to *the egg on the table*), or as a goal (*in the bowl over the flour* as opposed to *in the bowl beside the sugar*). Results showed that participants interpreted “in the bowl” as a modifier when there were two eggs in the visual environment that were liquid (i.e. the shell cracked open, with the contents uncooked and hence pourable); when only one egg was liquid, and the other was hard boiled, for example, listeners anticipated “in the bowl” to describe the goal.

Knoeferle et al. (2005) found, in an experiment investigating German thematic roles and grammatical relations, that visual information can influence thematic role assignment. For example, in a sentence like ”Die Prinzessin wäscht offensichtlich der Fechter,” (*The princess is apparently washing the fencer*, or *The fencer is obviously washing the princess*), “the princess” can ambiguously be assigned the role of agent and the relation of subject, or alternatively the role of patient and object. As neither role nor relation are overtly marked on nouns like “the princess” in German, resolution is only achieved when the final noun phrase, “der Fechter”, is encountered, as it is unambiguously nominative case, and hence functions as the subject. An accusative case designating object status would result in a different surface form, namely “den Fechter.” Fixation patterns suggested that listeners are able to use visual context immediately to assign thematic roles, and to disambiguate between interpretations.

1.1.1 Speaker-Related Variables

In addition to factors pertaining to the referential context, the speaker’s perceived identity – in essence, how someone looks or sounds, and how someone that looks and sounds a certain way is “supposed” to speak (see also the section on stereotyping below) – has been found to influence language comprehension processes as well.

In an early off-line ratings study, Rubin (1992) found that photographs, or rather the perceived ethnicity of the speaker *inferred* from those photographs, influenced how strong a foreign accent was perceived to be. Similarly, Woumans et al. (2015) suggest that faces can "prime" a language, at least as long as the face is a reliable clue for the same language across trials: For example, if bilingual listeners were trained to associate a face with the Spanish language (as opposed to Catalan), participants responded *faster* in the congruent condition in the follow-up association task, that is, when they produced a Spanish verb associated with a Spanish noun (as opposed to a Catalan verb and noun) while the "Spanish face" was displayed. This finding was replicated in a noun-noun association task in the same study.

In an EEG study, Van Berkum et al. (2008) analyzed ERP responses to statements colliding with a speaker's perceived identity, such as an adult male announcing that he wished he looked like Britney Spears. They found that such statements, clashing with Dutch stereotypes based on age, class, or gender, reliably elicited an N400 component. This component is generally elicited by all content words, but is significantly larger in amplitude for items that are difficult to integrate into the preceding context (see e.g. Allen et al. 2003; Kutas and Federmeier 2007).

The speaker's individual speaking style has been found to also influence language processing. Investigating the effect of a speaker's accent on language comprehension, Hanulíková et al. (2012) compared ERP responses to errors in native-accented and foreign-accented Dutch speech. Semantic anomalies elicited an N400 component irrespective of the speaker's accent. On the other hand, a P600 component, which is generally associated with syntactic violations[2] was *only* observed in response to gender agreement errors in native-accented speech. These results suggest that native-speaker status influences the expectations the listener has: With foreign-accented speech, listeners appear to be including errors in the representation they have formed about the speaker, and are thus not surprised by gender agreement errors. This is further supported by the fact that this surprisal, indicated by a significantly different ERP component during the early trials, vanishes towards the later trials – presumably when listeners had updated their expectations of the speaker's language such that it included the occasional error, indicating adaptation over time. Grey and Van Hell (2017) also found differing ERP responses depending on the accent of the speaker and how familiar the listener

[2] Whether purely linguistic or also non-linguistic remains a matter of debate; cf. Coulson et al. 1998; Osterhout 1999.

was with the foreign accent. These results highlight the importance of assessing the listener's language background, such as their familiarity with the accent at hand, as these factors were found to modulate responses systematically.

Viebahn et al. (2017) found that listeners only experienced a P600 component when a syntactic error occurred in careful speech as opposed to casual speech. The authors conclude that syntactic processing appears to remain undisrupted when the error occurs in casual speech as the error is *consistent* with that speaking style. It thus appears that the listener "expects" this kind of error to occur in this particular environment – "listeners take information about the speaking style of a talker into account when processing the acoustic-phonetic information provided by the speech signal." (Viebahn et al. 2017, p.2) These findings are supported by those in Romero-Rivas et al. (2016), where the comprehension of words semantically related to an expected continuation was cognitively more demanding when the speaker had a foreign accent.

1.1.1.1 Stereotyping

Many of the inferences made about a speaker are rooted in stereotypes, which form part of the listener's world knowledge. Very generally, stereotypes are defined as sets of ideas, or connotative associations, about how a certain group of people looks and behaves, and why they look and behave the way they do (see e.g. Schaller and Neuberg 2012; Strand 1999).

The process of stereotyping then assigns those ideas and properties to an individual based on the social group they are perceived to belong to (see e.g. Quadflieg and Macrae 2011), effectively "tagging" them with how they should, or ought to, behave. In linguistic research, prior biases have been found to affect phonetic convergence; that is, when individuals have a positive impression of a group of people, they are more likely to imitate this group's dialect (Babel 2010).

Stereotyping is believed to facilitate stimulus processing in a complex world (see e.g. Hilton and Hippel 1996; Macrae and Bodenhausen 2000; Quadflieg and Macrae 2011; Strand 1999 for excellent overviews, also regarding the formation, maintenance, and change of stereotypes). Instead of having to expend significant amounts of cognitive energy and time to get to know a particular individual, they are grouped – or categorized – based on salient features. It is this membership in a particular group that then gives rise to expectations and beliefs about how an individual should, or does, commonly behave:

> “[P]erceivers initially classify others according to salient social categories (such as sex, race, and age). More specifically, [...], others are seen as random instances of generic social groups until a perceiver is willing to engage in individuated impression formation (i.e., in making sense of others as unique entities) – either because a specific target gains heightened relevance or because it consistently fails to confirm category-based expectations.”
>
> Quadflieg and Macrae (2011, p. 221)

Stereotyping provides a (cognitively) cost-effective way of navigating the world, even though there is potential for error and overgeneralization, and, of course, stigmatization and the perpetuation of inequality (Hilton and Hippel 1996; Quadflieg and Macrae 2011); it appears to be “cheaper,” in a cognitive sense, to repair a wrong categorization every once in a while, than having to assess in detail every individual one encounters along the way (Oakhill et al. 2005). Importantly for this dissertation, the extent to which stereotypes are activated by a stimulus seems to differ based on listener-related individual differences, such as how prejudicial the perceiver is generally (Quadflieg and Macrae 2011); we will discuss these variables in greater detail in Section 1.1.2.

1.1.1.2 Gender Stereotyping

Of special interest for this dissertation are stereotypes based on gender, or rather the *perceived* gender, of a speaker. As one of the “big three” (gender, sometimes referred to as *sex* in the literature; race; and age), gender is one of the most salient features that are inferred about someone early on (Quadflieg and Macrae 2011). For example, gender stereotypes associated with certain occupations seem to be activated immediately and automatically. They also seem to be beyond strategic control, and hard to suppress (Oakhill et al. 2005; Pyykkönen et al. 2010). Information on an individual’s gender seems to form part of the listener’s world knowledge (see e.g. Banaji and Hardin 1996; Carreiras et al. 1996; Osterhout et al. 1997) that is integrated rapidly during language comprehension (see e.g. Marrville 2017; Pyykkönen et al. 2010). Stereotypical gender information has been shown to become activated even when it is not necessary for coherence, and/or when it is not stated explicitly (Pyykkönen et al. 2010), and can even form a more salient clue than syntactic information (Molinaro et al. 2016).

1.1.1.3 Voice-Based Gender Stereotyping

For this dissertation, greater focus will be placed on the gender that a listener *infers* about the speaker solely based on their voice, as opposed to, for example, the connotative gender feature commonly associated with a particular profession. Much of the existing linguistic gender stereotyping literature has focused on the latter; however, there is a growing field in which voice-based inferences are being investigated in more detail.

Generally, voices with lower formant frequencies, a lower fundamental frequency, and more resonance are characterized as male, due to both physical characteristics and dominant cultural norms (Ko et al. 2006; Strand 1999). Expectations around how an individual of a certain gender is *supposed* to sound have been shown to alter the perception of phonemes, suggesting the existence of contextual effects even within low-level speech perception (Strand 1999). In more recent research, voice has been described as a person's "auditory face," as it gives the listener clues to biological characteristics, such as age, fitness, and gender (Belin et al. 2011), but also personality traits (Aronovitch 1976), attractiveness, and current emotions (for an overview, see Ko et al. 2006). Hanulíková and Carreiras (2015) found an early ERP negativity when the inferred gender of a speaker did not match the predicate of a sentence, suggesting that gender information inferred from the voice signal is integrated into language comprehension early on.

As such, the gender feature that is inferred from the voice signal, along with its associated stereotypes, presents a good testing ground for the interplay of speaker- and listener-related variables (which we will now discuss in detail below) in language comprehension.

1.1.2 Listener-Related Variables

In addition to information that a listener infers about the speaker, the listener's own internal state – how they feel at the time, what they believe to be right or moral, and what their personality traits are – seems to influence language comprehension as well; we will now discuss these influences in detail in the following sections.

1.1.2.1 Political Values

Van Berkum et al. (2009) analyzed ERP responses while participants were filling in a political survey. Compared to a baseline of statements that participants agreed with, statements clashing with an individual's value system and political values, such as "I think euthanasia is an acceptable..." when the participant opposed this practice, again elicited a distinctive ERP response just 200-250ms after the onset of the critical word, in addition to an N400 component. These results suggest that the *listener's* values and beliefs also play a role in the same early meaning-making processes.

In addition, Marrville (2017) found, in an implicit causality experiment, that political values significantly modulated the effect that verb valence and dominance have on the implicit causality bias. Implicit causality verbs, such as "praise" or "apologize," bias participants as to which of the noun phrases in the sentence is the "cause" of an event. For example, "praise" biases the listener to assume that "Jack praised Jill because..." continues with "she" (the *bias-consistent* continuation) instead of "he" (the *bias-inconsistent* continuation), as the praisee's behaviour is assumed to be the cause for the praise. Conservative participants seemed to be affected differently by a verb's dominance and valence than their more progressive peers:

> "[W]hen the cause of the action was perceivable by progressive-leaning participants to be outside their control (i.e., low dominance) and had a positive effect (i.e., high valence), they attributed the action to the recipient of the event [...]. Conversely, when the low dominance action had a negative effect (i.e., low valence), they attributed cause to the character who was performing the action."
>
> Marrville (2017, p. 77)

Although not much is known about the intersection between political values and language comprehension, political values have been of interest in different areas of general cognition, some of which we will return to in Section 1.1.2.4 below. The four experiments that will be reported below thus present a significant addition to the research on political values in relation to language comprehension.

1.1.2.2 Mood

Mood, a more transitory state than moral beliefs, was also found to affect language comprehension in an implicit causality experiment (Van Berkum et al. 2013): A good mood caused listeners to engage in more anticipation as to what the referent might be. This was reflected in a distinctive ERP component in response to a bias-inconsistent continuation ("he" in our example of "Jack praised Jill because..."); a bad mood, on the other hand, effectively stifled any anticipation.[3]

Even a simulated mood appears to affect processing speed, such that processing is faster when an individual's facial expression matched the valence of the sentence. For example, a "pleasant" sentence, in which you are called to the stage and praised for an achievement, was read faster when the facial expression approximated a smile (with a pen held sideways between the teeth); "unpleasant" sentences, such as one informing you that the police pulls you over for an infraction, were read faster when the facial expression approximated a frown (pen held between lips, sticking outward; Havas et al. 2007; see also Havas et al. 2010). Mood, as a decidedly extra-linguistic factor pertaining to the listener, hence appears to influence language comprehension from an early stage as well.

1.1.2.3 Personality Traits

Not much is known about the influence of a listener's personality on automated language comprehension. However, an individual's personality has been known to significantly influence important aspects of one's life, such as academic motivation, academic success, the choice of learning style (Busato et al. 1998, 2000; Chamorro-Premuzic and Furnham 2009; De Raad and Schouwenburg 1996; Furnham et al. 1999; Gill and Oberlander 2002; Jensen 2015; Komarraju and Karau 2005), work performance (De Raad and Schouwenburg 1996; Furnham et al. 1999), second language acquisition (Robinson et al. 1994), information-seeking behaviour (Heinström 2005), the choice of romantic partners and friends (Wu et al. 2017), and the use of online social media (Moore and McElroy 2012; Wehrli 2008).

[3] In this context, also note the discussion surrounding the terminology around prediction and anticipation in Section 1.2.2.

Additionally, there is ample evidence suggesting that an individual's personality significantly influences their natural written language use (Gill and Oberlander 2002, 2003; Oberlander and Gill 2004): For example, highly extraverted[4][5] individuals seem to use more and looser punctuation marks, more assertive language, and higher rates of plural *we*, such that extraversion can be detected reliably in email communication (Gill and Oberlander 2002, 2003). As of 2019, there is now at least one texting app that analyzes one's conversation partner along several personality dimensions, and gives recommendations on how to communicate with them effectively (Wiggers 2018).

Boland and Queen (2016) have found, in an off-line ratings study, that a reader's reaction to errors in response to a supposed housing advertisement was significantly influenced by their personality traits. Readers were presented with email replies that contained typographical errors ("typos," such as *teh* for *the*) and homophonous grammatical errors (termed "grammos," such as *to* for *too*, or *there* for *their*, and vice versa). Less agreeable readers were found to be influenced more negatively by "grammos;" the same was found for more conscientious and less open people in response to "typos." Generally, more extraverted readers tended more towards overlooking errors, whereas more introverted readers had a more judgmental approach. The level of extraversion was also found to influence speech production in both native speakers and second language learners, where more extraverted learners were found to be more fluent than their introverted counterparts (Dewaele and Furnham 1999, 2000). Social media posts have been found to be appropriate as a language-based personality assessment (Park et al. 2015), so that linguistic style is now considered a reliable individual difference, correlated with an individual's personality (Pennebaker and King 1999; Pennebaker et al. 2003).

One of the basic notions of personality, namely where an individual is situated on the introversion/extraversion scale, is believed to have a physiological basis – it is assumed that an organism generally tries to be within optimal levels of arousal: While an introvert is gen-

[4] Note that two different spellings seem to be in use, both for the noun referring to the trait (*extraversion* vs. *extroversion*), and for the corresponding adjective (*extraverted* vs. *extroverted*). Interestingly, today, *extraversion* and *extroverted* seem to be the more common variants, with *extroversion* and *extraverted* trailing behind (Michel et al. 2011). For consistency, I have adopted the spellings of *extraversion* and *extraverted* in this dissertation.

[5] The personality traits in this section, such as Agreeableness or Conscientiousness, describe the traits assessed in the *Big Five* personality battery that was also used in all four experiments in this dissertation. For details on this test, please see John and Srivastava (1999) and John et al. (1991), and refer to Section 2.4 and Table 2.2.

erally already at the optimal arousal level, or even exceeds it, without external stimulation, an extravert's arousal level, as measured through, for example, skin conductance or brain waves (Eysenck 1990 as cited in Jang 1998), is assumed to rest below the optimal threshold, so that they seek external stimulation to reach the optimal level (Dewaele and Furnham 1999). Patterns of cortical blood flow further support the idea that there is a physiological basis for personality traits – Stenberg et al. (1990) found that there was higher activity in the temporal lobes for introverts. In addition, there appears to be a neurological basis for the similarities – and differences – between self- and other-representation, a concept crucial for empathetic reactions (Jackson et al. 2006). As one specific personality trait that has received some attention in language comprehension research, empathy has been shown to influence ERP responses to unusual stimuli: Re-using the materials in Van Berkum et al. (2008), but adding the listener's empathy level assessed via the *Empathy Questionnaire* (*EQ*; Baron-Cohen et al. 2001) as an additional variable, Van den Brink et al. (2010) found that listeners with high empathy levels showed a significantly larger N400 component in response to socially contradictory information than those participants with low empathy scores. So, for example, a high-empathizing listener would experience a significantly larger N400 when hearing a child say "Every evening I drink some wine before I go to sleep" than a listener with low empathy scores. Similarly to the effects of a good mood in Van Berkum et al. (2013), which was discussed in Section 1.1.2.2, the ability to empathize to a higher degree is assumed to encourage more anticipation based on inferences about the speaker – such as their age, and what is an appropriate beverage to consume at the inferred age, in our example above – and hence experiencing surprisal at an unexpected item. Conversely, individuals with lower empathy scores seem to rely less on contextual cues. These results are corroborated by an fMRI study on Mandarin Chinese by Li et al. (2014), where participants listened to unusual utterances that were created by inserting a highly common or expected event into a construction that is used to highlight unexpected or unusual events (similar to English *even*). An individual's empathy levels, operationalized as scores on the *Fantasy* and *Perspective-Taking* subscales of the *Interpersonal Reactivity Index* (*IRI*; Davis 1980), were shown to modulate brain activity in different regions related to the processing of utterances where a common or expected event followed a construction that is only used with unlikely or unexpected events.

Summing up, it is assumed that a good mood, or high empathy levels, “[tell] you to trust your instincts and go out there to explore” (Van Berkum et al. 2013, p. 2). A proposed tie-in with emotion processing comes from Havas et al. (2007), who relate their results to theories in which emotions are assumed to change affordances, the links between perception and action (Jackson and Decety 2004): In this view, a positive mood prepares the body to approach, whereas a negative mood prepares the body to avoid. The findings in Van Berkum et al. (2013) for mood, and in Van den Brink et al. (2010) for empathy, suggest that both mood and empathy levels might have an effect on affordances, and hence change how strongly a human engages in “approaching” or “exploring”, or how much they stay put and rely on bottom-up information. A related take on this notion can be found in the *bio-energetic account*, which suggests that emotional states signal the amount of cognitive resources available for more “costly” behaviours, such as exploration and anticipation (Van Berkum et al. 2013; Zadra and Clore 2011). Note that this is tightly linked to the notion of anticipation in linguistic processing, which will be discussed in Section 1.2.2.

1.1.2.4 Disgust Sensitivity & The Behavioural Immune System

Another, rather special, kind of information that may form part of the listener’s world knowledge in language comprehension is Disgust Sensitivity – or, more broadly, how much an individual tries to keep harmful pathogens at bay. To understand how this variable may affect language comprehension, this section presents a short foray into general cognition – much of which is not taking place within the realm of conscious thought. Sub-conscious response mechanisms, such as those in place to avoid threats, developed a long time before conscious decision-making, and are still constantly “monitoring” the environment (Aarøe et al. 2017; Bargh and Chartrand 1999). The brain is thus thought to have been shaped, by selectional pressures over millions of years, not to provide exact calculations, but rather to form fast, adaptive responses in situations that might be hazardous to fitness and/or reproduction (Neuberg et al. 2011). Even small amounts of self-control seem to use up the limited self-regulatory resources, so that the conscious decision-making system can only be used occasionally, and most cognition hence relies on sub-conscious thought (Bargh and Chartrand 1999; Baumeister et al. 1998). Basic emotions, such as anger and fear, are commonly thought to be sub-conscious triggers for behaviour that attempts to redirect attention to the threat in question, with the goal to instigate the actions necessary to mitigate or evade the threat (Cottrell and Neuberg 2005; Neuberg et al. 2011; Schaller and Neuberg 2012). For example, when faced with a dangerous predator, it follows logically that contact is to be avoided to ensure survival; this can be easily learned and taught to offspring, ensuring that this behavioural response became a crucial part of the human subconscious threat response repertoire (Murray and Schaller 2016). However, not all threats to survival and reproduction are as obvious as a charging predator, or a violent thunderstorm from which best to seek shelter.

While human populations developed ultra-social behaviour, including, for example, forming close-knit groups, to better protect themselves from threats, those same close-knit groups can be detrimental – or even disastrous – in terms of disease and pathogen transmission (Murray and Schaller 2016; Schaller and Neuberg 2012). A close-knit group might offer better protection from a tiger, but it also facilitates pathogen transmission – especially in pre-medical times, where effective countermeasures, details of pathogen transmission, or cures to common diseases were unknown. Pathogens, such as bacteria and viruses, differ from other, more obvious threats in that they are invisible. Unlike "threats from without," pathogens can be perceived only indirectly, e.g. through foul smells or an appearance that is "off" (Aunger and Curtis 2013; Murray and Schaller 2016; Tybur et al. 2013). However, the two types of threats are alike in that they have exerted similar selection pressures over human populations for a long time, and in that they are equally lethal (Aarøe et al. 2017; Murray and Schaller 2016). A "threat from within" thus requires a response as well and cannot simply go untreated – but it requires a distinctly *different* response than a highly visible threat. Significant differences in autonomous nervous system responses, as measured through e.g. heart rate and respiration rate, further support the notion that the response to a (perceived) pathogen threat is distinct from the feeling of fear, both in terms of physical responses and threat perception (Murray and Schaller 2016).

The need for a response to pathogens likely resulted in the formation of the highly effective human immune system, which triggers appropriate responses once it detects that a pathogen has entered the body (Schaller and Neuberg 2012). However, while an immunological response is highly effective, it is also extremely costly – responses such as inflammation and fever, which both raise body temperature, either locally or globally, have a high energetic cost (Murray and Schaller 2016) and can prevent the human from spending their time on other behaviours that support their existence and well-being, such as looking for sustenance, or mating (Murray and Schaller 2016; Neuberg et al. 2011). Additionally, in the time that it takes for the immune system to formulate its responses, the pathogen can already do significant damage. A preventative mechanism is hence desirable, so that the immune system only has to kick in and consume resources as a "last resort" (Murray and Schaller 2016). Despite, or because of, pathogens' invisible nature, humans have learned to associate certain cues with the presence of pathogens over time (Faulkner et al. 2004). However, it was not until quite recently in human history that the transmission of pathogens has been researched and understood sufficiently to limit their spread through medical means, or to develop successful medical responses for when pathogens had already entered the body and had begun to cause harm. For the majority of their existence, human populations have indeed *not* been able to rely on relatively modern inventions such as public health advisories regarding hygiene, vaccinations, or even detailed knowledge on how different diseases are actually transmitted. As such, the most effective protection from pathogens that was available *prior* to the developments of modern medicine, i.e. for most of human history, was the feeling of disgust, the "affective signature" (Murray and Schaller 2016, p. 115) of a response to objects or individuals that are commonly associated with posing a risk of pathogen infection. Disgust, as such, is now thought to be closely associated with the *Behavioural Immune System (BIS).* While other theories of disgust have been proposed over the years (for details, see McGinn 2011), they are largely dated and/or have been criticized as not being able to account for recent empirical evidence (Strohminger 2014). For example, the *Taste-Toxicity Theory* relates the feeling of disgust to the presence of toxins, while the *Animal-Heritage Theory* proposes that disgust is elicited by items and situations that remind us of our animal status. The latter ignores the threat posed by pathogens and the fact that revolting tastes or smells elicit disgust, while neither of the two theories can account for the widespread influence that disgust seems to have on the perception of out-groups, and the avoidant behaviour (which will be discussed below) that results from it.

The BIS is thus the most widely accepted comprehensive theory of disgust today (Strohminger 2014), and, for the purposes of this dissertation, I will adopt this approach.

The Behavioural Immune System

Unlike the regular human immune system, which responds once pathogens are already *in* the body, the BIS tries to prevent pathogens from *entering* an organism in the first place (Aarøe et al. 2017; Murray and Schaller 2016; Neuberg et al. 2011; Schaller and Park 2011). Very broadly, it does so by scanning the environment for potential pathogen threats, and, once it detects a potential threat, motivates actions and responses that help avoid pathogen contamination (Aarøe et al. 2017; Neuberg et al. 2011; Schaller and Park 2011).

The feeling of disgust is thought to be the "affective signature" of BIS activity (Murray and Schaller 2016, p. 115), and is triggered by a (perceived) immediate threat of infection, directing attention to the threat with the goal to avoid contact with the threatening item – such as rotten food, feces, or indeed individuals that are assumed to be carriers of an infectious disease (Faulkner et al. 2004; Murray and Schaller 2016; Oaten et al. 2009). Disgust is considered a highly visceral and basic response: Chapman et al. (2013) found that participants showed more recall and recognition for disgusting stimuli, even when they were exactly as arousing as fear-inducing stimuli, suggesting that disgust has special significance within human cognition, and that it is central to survival (Inbar et al. 2011). The feeling of disgust and its associated responses and behaviours likely developed from a direct physical response to ingesting poisonous or rotten food, as indicated via a foul taste or smell (Faulkner et al. 2004; Murray and Schaller 2016). The feeling of disgust would protect the individual either by giving clues that this food was best avoided, or to expel it through vomiting if it had already been ingested (Haidt et al. 1994). As many diseases cause the prototypical human appearance to change, for example through skin rashes, discolouring, sneezing, or coughing, the system has adapted to detect *any* deviation from the prototypical human appearance (Schaller and Neuberg 2012). These "deviations" can, and often will, include the default appearances of "other", unfamiliar groups of humans, as such "outgroups" historically had the potential to carry different, novel pathogens, thus posing an even greater risk than familiar pathogens, as antibodies may not be present in the local, "familiar" population (Aarøe et al. 2017; Faulkner et al. 2004). As will be discussed in greater detail below, this is thought to be tightly connected to racial prejudice.

However, it is not only deviations in physical appearance that have been found to trigger a BIS response, but also differences in *behaviour* that may have become associated with increased pathogen presence, such as different habits regarding hygiene or cooking. As such, it can be both physical *and* cultural differences that may trigger disgust and BIS activity, even if no pathogens are present (Aarøe et al. 2017). Differences in appearance or behaviour thus seem to function as "informational tags" regarding an individual's health or their membership in an outgroup, even though the outgroup may not actually be transmitters of disease (Schaller and Neuberg 2012; see also the *Smoke Detector Principle* discussed below). This has important implications for ingroup/outgroup prejudice, stigmatization, and moral judgment (Aarøe et al. 2017; Faulkner et al. 2004; Murray and Schaller 2016), as we will return to below, and as is of special importance for this dissertation.

While avoidance behaviour can limit a human's exposure to pathogens to almost zero, such as is the motivation behind quarantine for patients with immunodeficiency disorders, it also limits the possibility of contact with other humans. This may sound like desirable behaviour if it succeeds in preventing illness – but it carries a cost: As an ultra-social species, avoiding other humans can mean forming fewer alliances, and hence a reduced chance of survival and procreation (Aarøe et al. 2017). Accordingly, as continuous, unmitigated exposure to pathogens can be detrimental, and full avoidance behaviour in all situations comes at a great cost as well, the BIS is thought to be finely calibrated to calculate the benefits versus the cost of avoidance behaviour in the current situation, taking into account environmental variables and information indicating how vulnerable to infection the individual is at the time. As a result, the BIS is adjustable in strength, and its sensitivity differs across individuals – known as the **Functional Flexibility Principle**: When vulnerability is salient, the benefits of pathogen avoidance outweigh the costs, and the BIS responds more strongly; when the environment and the perception of it suggest relative protection from pathogen threat, the BIS retreats and responds less strongly (Murray and Schaller 2016; Schaller and Neuberg 2012; Schaller and Park 2011).

It should be noted that it is not just abstract and objective environmental variables that factor in this calculation. The cost-to-benefit ratio, or attentiveness to pathogen cues, is modulated by how vulnerable to infection an individual *perceives themselves to be* (Schaller and Neuberg 2012; Schaller and Park 2011). This feeling of pathogen vulnerability varies within an individual, such as when the presence of disease is made salient versus when temporary protection from pathogens is available (e.g. Faulkner et al. 2004), and *between* individuals, simply depending on how much a person believes themselves to be vulnerable to infection and disease generally. This relates strongly to individual differences in Disgust Sensitivity mentioned previously (Haidt et al. 1994; Inbar et al. 2009, 2011).

Like all human threat management and threat avoidance systems, the BIS is calibrated to infer threat when cues only *imply* threat (Neuberg et al. 2011). To infer threat, the BIS relies on superficial cues, which are far from perfect indicators of pathogen presence, to compute the probability of pathogen presence (Murray and Schaller 2016; Schaller and Neuberg 2012; Schaller and Park 2011; Tybur et al. 2013). In the computation of this likelihood, the system is commonly likened to a **smoke detector**. For any signal detection mechanism, such as a smoke detector or the BIS, two errors are possible: A false positive, where a threat/fire is detected when there is none; this is irritating or unnecessary, but generally not fatal. The other kind of error is a false negative, where an actual threat/fire remains undetected. This can be less irritating at the time, but has the potential to be extremely costly in terms of health or continued existence. Like a smoke detector, the BIS is hence calibrated to "overgeneralize" and give lots of false positives, rather than risking to miss an actual threat (Murray and Schaller 2016; Schaller and Neuberg 2012; Schaller and Park 2011; Tybur et al. 2013).

Outgroup Stigmatization and Prejudice

The hypervigilance of the BIS has important implications for the understanding of stereotypes, and the related prejudicial and stigmatizing behaviour. This extends even to concepts that are only very tangentially related, or even entirely unrelated, to pathogen threat, such as immigration: "[P]hysical as well as cultural differences may be mentally tagged by the behavioural immune system as signs of pathogen risk, eliciting disgust, and causing people to avoid contact with ethnically different individuals and prefer restrictive immigration policies" (Aarøe et al. 2017, p.278). While mental categories are a necessary component of human cognition, needed to process the general overabundance of information in our surroundings quickly, stereotypes are considered maladaptive forms of mental categories, as they do not correspond accurately to the evidence on hand (Bargh and Chartrand 1999; also recall our earlier discussion of stereotypes in Section 1.1.1.1.) *Prejudices* are commonly defined as specific feelings towards a group of individuals that result in unequal, typically less fair, treatment simply because of the group's categorical characteristics (Schaller and Neuberg 2012). They are thought to be an "adaptive consequence to a distinct kind of threat that imposed evolutionary selection pressures on ancestral populations"(Schaller and Neuberg 2012, p.4). Many historical threats that prejudices are thought to have originated from could be inferred from aspects of physical appearance, such as rashes, as discussed earlier. However, the BIS regards *any* deviation from the "prototype" as a cue for threat, including disfigurements due to a disability, or even a difference in skin colour or weight. Another set of features that can be cues for BIS activity and avoidance behaviour are customs, such as those regarding food preparation or personal hygiene. Before modern medical interventions, such as immunization and antibiotics, became commonplace, and before humanity learned how exactly pathogens work and are transmitted, cultural norms regarding food preparation and hygiene likely were "safeguards" against pathogen transmission (Murray and Schaller 2016; Schaller and Neuberg 2012). Cues for outgroup status can hence be outward appearance, such as for persons with disabilities or indeed those of a different skin colour, or behavioural customs, such as for immigrant groups or homosexual individuals. Both sets of features – those relating to appearance and those relating to customs – can, and often will, result in ethnocentric and xenophobic responses (Cottrell and Neuberg 2005; Murray and Schaller 2016; Schaller and Neuberg 2012; Tybur et al. 2013). Xenophobic propaganda often alludes to this pathogen-related origin of prejudicial behaviour, likening ethnic outgroups to rats, flies, and lice, which are typically associated with transmitting disease (Schaller and Neuberg 2012). In this context, it should be noted that the feeling of disgust has been described as reflecting an *essentialist* view of a group; that is, if someone feels disgust toward a particular out-group, they are more likely to believe that the out-group is different from the in-group due to differences in biology, and that the group has defining features and membership in it is fixed (Katzir et al. 2018).

As per the Functional Flexibility Principle, discussed above, BIS activity and sensitivity to stimuli are not the same across the board: Stigmatization and prejudicial, discriminatory behaviour is increased in situations where an individual feels more vulnerable to or threatened by pathogens (Faulkner et al. 2004; Murray and Schaller 2016; Oaten et al. 2011; Schaller and Neuberg 2012). As discussed previously, differences in vulnerability differ *between* individuals, where people with greater concern about pathogens, or with higher

Disgust Sensitivity, generally show more prejudicial reactions against individuals perceived to be a threat based on outgroup status; and *within* individuals, such as when a threat is made salient. For example, the immigration of foreign-looking immigrants was seen as less favourable, and policies that would favour *familiar* immigrants were endorsed more, in a disease-salient condition (Faulkner et al. 2004). Additionally, when a threat was made salient, “ambiguous” members of a group were found to be categorized as members of the outgroup, and individuals were found to especially value conformity and obedient behaviour (Murray and Schaller 2016).[6] This suggests that conformity to norms has likely played an important role within ancestral populations in mitigating infection risk; a trait commonly associated with authority/structure and a conservative political leaning (Graham et al. 2009; Haidt and Graham 2007; Jost et al. 2003).

Political Views & Disgust

As discussed in detail in the previous section, disgust and BIS activity contribute to shaping the perception of (out-)groups, where “subjective foreign-ness” triggers avoidance (Faulkner et al. 2004; Inbar et al. 2009). There is now a growing body of research suggesting that this perception is not happening in isolation from other aspects of human cognition. In fact, “many worldwide cultural differences – in personality, values, and behavior – may be partially the product of psychological responses to the threat of infection” (Murray and Schaller 2016, p. 109). Of special interest for this dissertation are correlations that have been attested between Disgust Sensitivity and political values on the one hand, and Disgust Sensitivity and personality traits on the other, especially as there is little (or no, in the case of disgust) research on their influence on language comprehension.

It is important to note here that, for the purposes of this dissertation, the terms *progressive* and *conservative* are not to be taken to refer to distinctions or party affiliations in current politics (such as, for example, Democrats vs. Republicans in the United States, or supporters of the Liberal Party vs. those of the Conservative or Progressive Conservative Party of Canada). Rather, the terms are to be understood as two opposite ends of a continuum on which an individual may be situated regarding their values and beliefs. While the term “progressive” will be used throughout this dissertation to refer to the non-conservative

[6] A more detailed summary of several studies that link threat perception to outgroup bias and prejudices can be found in Neuberg et al. (2011).

end of the continuum, note that some sources refer to a distinction between *liberal* and conservative views; in those cases, the original wording will be retained. According to the political literature, individuals on the conservative end of the continuum generally value "maintenance of and conformity to traditional social norms" (Murray and Schaller 2016, p.103). Some of the most crucial differences between conservatives and progressives thus are differences in resistance to change; in attitudes towards equality, fairness, and authority; and in tolerance of ambiguity (see e.g. Graham et al. 2009; Haidt and Graham 2007; Jost et al. 2003). Whereas conservatives tend to resist change, to not regard equality as an important goal, value authority, and tend to not respond well to ambiguous stimuli, progressives overall tend to embrace change, view equality and fairness as an important goal of modern (in a temporal sense, not as a qualitative judgment) human societies, and are tolerant of ambiguity in life (Jost et al. 2003). This relates strongly to the different moral "foundations" that people on both ends of the political spectrum tend to base their judgments, voting behaviour, and and decision-making on. Whereas more progressive individuals appear to base their decisions on mostly two moral foundations, namely harm avoidance and fairness/equity, conservatives appear to, in addition, factor in purity, in-group loyalty, and authority/structure (Graham et al. 2009; Haidt and Graham 2007; Inbar et al. 2011). The purity factor is, as discussed previously, strongly related to disgust – and presents a strong link between politics and disgust, as conservatives see the "maintenance of purity as an inherent moral good" (Inbar et al. 2009, p. 715). As such, BIS activity and outgroup stigmatization extend to moral judgment: Norm violations are judged more harshly by people who are more easily disgusted (Murray and Schaller 2016). Moral disgust, which denotes the feeling of disgust that is triggered when one is faced with an immoral act, even if the act in question has nothing to do with pathogen presence or threat, appears to be rooted in the actual literal feeling of disgust – it is associated with movement in the same muscle regions that bad tastes elicit (Chapman et al. 2009; Tybur et al. 2013). Consequently, individuals who are relatively more prone to experiencing disgust have been found to be more likely to be conservative (Murray and Schaller 2016), especially regarding issues that relate to purity. Purity-related issues include, for example, gay marriage, or immigration from foreign lands (Inbar et al. 2009; Smith et al. 2011). These findings suggest that an individual high in Disgust Sensitivity does not oppose immigration because they are conservative, but rather because of their BIS responding more strongly to the idea of increased immigration of foreign people, in an attempt to manage the perceived pathogen threat (Aarøe et al. 2017). Additionally, Ahn et al. (2014) found in an fMRI study that progressive-leaning and conservative-leaning individuals exhibited different patterns of brain activity when viewing disgusting stimuli.

However, it is not only an individual's static, "baseline" Disgust Sensitivity that influences political feelings; much like outgroup stigmatization and BIS activity is increased in situations where a perceived pathogen threat is made more salient, momentarily inducing disgust, or making disease salient, also shifts individual political perspectives towards conservativism (Faulkner et al. 2004; Inbar et al. 2011). For example, Helzer and Pizarro (2011) found that, when participants were reminded to keep physically clean, they exercised harsher moral judgments regarding sexual purity, and reported being more politically conservative.[7] In addition, individuals tend to vote for attractive candidates in political elections, i.e. those candidates that conform to the default morphological appearance of the in-group, and that do not have "foreign" blemishes. Again, this behavioural pattern was found to become even stronger when a disease threat was salient (White et al. 2013).

Summing up, as Murray and Schaller (2016, p. 112) put it, "conformity, conservatism, and moral judgments are all motivated, in part, by the psychology of disease avoidance" – in addition to situational and socio-demographic factors, like the stability of the current political system, the individual's geographical location, and their education level (Inbar et al. 2009).

Fewer links have been observed between personality traits and Disgust Sensitivity; however, Druschel and Sherman (1999) and Haidt et al. (1994) found that Neuroticism[8] was correlated positively with Disgust Sensitivity (.45 and .23 respectively), and Druschel and Sherman (1999) in addition found an inverse correlation between Openness and Disgust Sensitivity (-.28). In Experiment IV (see Section 5), where Disgust Sensitivity was assessed via a post-test, the strongest correlation was found between Agreeableness and Disgust Sensitivity (.40; see Table 1.2), an interaction also attested in Druschel and Sherman (1999), who reason that "the disgust sensitive individual is likely to display characteristics of altruism, sympathy, co-operation and sensitivity to interpersonal needs of others" (p. 746).

[7] Going beyond the scope of this dissertation, this means that public health advisories could potentially have unintended political consequences (Murray and Schaller 2016).

[8] As noted previously, personality traits in this section, such as Neuroticism or Openness, again describe the traits assessed in the *Big Five* personality battery that was also used in all four experiments in this dissertation. For details on this test, please see John and Srivastava (1999) and John et al. (1991) and refer to Section 2.4 and Table 2.2.

The second-strongest correlation, at .29, was observed between Neuroticism and Disgust Sensitivity, thus corroborating the findings in Druschel and Sherman (1999) and Haidt et al. (1994). Openness was only very weakly negatively related, a correlation considered insignificant (-.09; p = .43). In the literature, Openness was also found to be related to tolerance of ambiguity, one of they key differences between conservatives and progressives (Jost et al. 2003). While these correlations and associations are interesting factoids, they serve merely as an interesting backdrop for this dissertation, but are not of immediate importance to the research questions.

Summing up, a number of subconscious factors seem to be relevant in social decision-making processes. Especially Disgust Sensitivity and the perception of pathogen threat, along with its related goal of disease avoidance, seem to be affecting such seemingly unrelated matters as an individual's political values and their attitudes towards immigration. Importantly for this dissertation, and especially for Experiment IV, these "seemingly unrelated matters" include attitudes towards "foreign" appearances and customs – both strong bases for prejudicial and stigmatizing behaviour towards foreign outgroups. For this reason, the *Disgust Scale – Revised* (*DS-R*; Haidt, McCauley & Rozin, 1994, modified by Olatunji et al. 2007) was administered to participants in Experiment IV, which – like all experiments in this dissertation – presented participants with socio-cultural clashes in addition to morphosyntactic and semantic anomalies. The findings summarized above thus run counter the idea of political thought originating solely from conscious thought (Inbar et al. 2011). Of course, such a proposition may – understandably – be received as controversial; by no means is the above research to be interpreted as suggesting that political choices and prejudicial behaviour are fully determined by nature, and not at all subject to subconscious thought. Instead, the research proposes that *some* aspects of human psychology, which likely originated in ancestral populations, giving them a leg up in the game of survival, have now been found to affect political values and attitudes *to a certain extent.* As Schaller and Neuberg (2012, p. 46) put it:

> "There are a variety of reasons that contribute to wariness about evolutionary approaches to human behavior [...] Among these reasons, perhaps, is our distinctly human fondness for the distinctively human wonders of cognitive rationality, which may lead people to reflexively recoil from the ugliness of our bestial

> past. But most prejudices are not cognitively rational products of our newfangled neocortex. Prejudices are, and always have been, products of the more ancient and beastly parts of our brains. If we ignore our evolutionary past, we are likely to ignorantly fall prey to the prejudices that have resulted from it. If we confront our evolutionary past (and its psychological consequences) with scholarly rigor, we can more truly know the nature of these prejudices and do something about them."

This dissertation presents one of the first analyses of the influence of this evolutionary past, as instrumentalized through Disgust Sensitivity, on language comprehension processes.

1.2 Anticipation

As already mentioned at the beginning of this Introduction, it is now generally assumed that language comprehension operates in a one-step fashion, with non-linguistic context being considered immediately, affecting anticipation regarding what the upcoming segment may be. Thus, linguistic anticipation is not a notion separate from the influences of extra-linguistic information, such as those discussed previously; the two are intertwined in that information, whether intra- or extra-linguistic, can become a factor in linguistic anticipation. In this section, anticipation will be discussed in regards to both general cognition and linguistic processing, as the latter will be assumed to be tightly related to general cognition for the purposes of this dissertation (Bybee 2010; Goodman and Stuhlmüller 2013; Van Boxtel and Bocker 2004; see also Marrville 2017 and the discussion in Chapter 6) – that is, language is assumed to recruit the same neural networks that are in use for general cognition, and heuristics and skills can be shared between the two.

1.2.1 Anticipation in General Cognition

Anticipation is considered an integral part of human everyday behaviour and various cognitive functions, such as vision, learning, causality, probability, planning, and more (Pezzulo et al. 2007; Riegler 2001; Tressoldi 2015; Van Berkum et al. 2005; Van Boxtel and Bocker 2004). It makes preparation for a situation possible – cognition without anticipation would render all behaviour "exclusively reactive" (Van Boxtel and Bocker 2004, p. 61). Even seemingly simple actions, such as turning on a faucet to get water, or picking up a pen to start writing, involve a great deal of subconscious anticipation about how the world works: We anticipate that water will flow if we turn on the tap, that the ink contained in the pen will stick to the writing surface, and so on and so forth (for more examples, see Riegler 2001). However, anticipation does not simply facilitate the human existence, or cause actions to be more efficient than they would be without anticipatory powers; instead, it is necessary for survival (Riegler 2001). The anticipation of a sensation, for example a feeling of heat or pain, is crucial in planning an action that will avoid the sensation –and the related threat – in question (Pezzulo et al. 2007). As such, anticipation is a crucial component of how the subconscious brain operates; Pezzulo et al. (2007) describe anticipation as being "at the core of cognition" (p. 68). In daily life, humans generally anticipate that "similar issues have similar causes" (Riegler 2001, p. 535): Anticipating that a solution that has worked before will work again is an important contributor to not getting bogged down by the combinatorial explosion of choices whenever a decision has to be made (Riegler 2001, p. 535). Based on the results reported in Kutas (1997), Van Boxtel and Bocker (2004) explicitly consider anticipation in linguistic processing to be one application of cognitive anticipation, rather than language being modular, domain-specific, and entirely separate from general cognition. Further support for language being closely intertwined with general cognition comes from experimental research in which several intra- and extra-linguistic variables that affect language processing, some of which were discussed in detail in Sections 1.1.1 and 1.1.2, have also been found to affect general cognition. For example, recent research suggests that mood has a similar effect on both general cognition and on language processing (recall our earlier discussion in Section 1.1.2.2): Mood can change perception – a good mood literally makes a hill look less steep, and generally encourages individuals to rely on heuristics rather than detailed information. A bad mood, on the other hand, seems to be associated with a more skeptical approach, waiting for more information to come in, and watching for details (Zadra and Clore 2011). Results in Van Berkum et al. (2013) support this notion, where an induced bad mood indeed made listeners rely less on heuristics, hence engage less in anticipation, and attend more to the bottom-up signal than did those listeners in a good mood. Also recall our earlier discussion of the effects of a simulated mood in Section 1.1.2.2 (cf. Havas et al. 2010, 2007). Further evidence comes from Tanenhaus et al. (1995), where visual input modulated syntactic processing; if language and cognition were not intertwined, visual input should not mediate linguistic processing this rapidly. In fact, many other contextual variables – such as personality traits and political views, as discussed in Section 1.1.2 – have been found to affect language comprehension, further linking aspects of non-linguistic cognition to language comprehension. We will now turn to a discussion of anticipation as it relates specifically to language comprehension.

1.2.2 Anticipation in Language Comprehension

The notion of linguistic anticipation is tightly linked to the immediate integration of contextual information, as already discussed in Section 1.1. If the syntax of a given utterance is considered to be the only constraint in the first step of comprehension, as is the case in the standard two-step model of language interpretation (see e.g. Chomsky 1957; Cutler and Clifton 1999; Grice 1975), then there are thousands of words available that could conceivably follow any given segment, each at a very low cloze probability. Anticipatory processes would seldom be successful, and would thus be inefficient and hardly useful (Van Berkum et al. 2005).

However, a growing body of research suggests that it is *not* just syntax that affects the early stage of linguistic interpretation, but rather syntax *in addition to* many different types of contextual information – from visual context to prior discourse, and even co-speech gestures, so that meaning is contextualized from the start, and not integrated with the wider discourse context in a second step (Federmeier 2007; Hagoort et al. 2004; Hagoort and Van Berkum 2007; Levy 2008; Traxler 2014; Van Berkum et al. 2005, 2008; Van Petten and Luka 2012): "In all, the recent evidence converges to suggest that, when comprehending sufficiently constraining yet natural fragments of discourse, listeners and readers do anticipate upcoming words on the fly as the text unfolds" (Otten and Van Berkum 2008, p.467; also see Van Berkum et al. 2005, p. 460, for a highly similar notion). This shift away from a two-step model, where comprehension is assumed to proceed first based only on the utterance itself, to a unified approach where contextual information influences processing from the start, was spurred by the fact that a two-step model was not able to sufficiently explain attentional shifts to items before they have been encountered, and the speed with which humans interpret language (Altmann and Kamide 1999; Traxler 2014). In other words, discourse-based anticipation[9] is possible, and parsimonious, only in a framework in which contextual information is considered at the same time as syntactic constraints, as it is this same contextual information that constrains upcoming choices sufficiently for anticipation to be feasible and useful. It should be noted that the extra-linguistic factors discussed in Section 1.1 are not to be considered separately from anticipatory processes, or from one-step modelling. Rather, these factors are a part of the contextual information that is considered along with syntactical constraints: For example, in Van den Brink et al. (2010), high empathizers were found to make more use of contextual cues than their low empathizing counterparts. That is, high empathizers did *not* compute meaning first on a syntax-only basis, to then relate it to the communicative context later; rather, they considered speaker information right from the start. Low empathizers, on the other hand, seemed to employ a more syntax-first approach. The debate has thus moved from whether there is any anticipatory processing at all to the more nuanced aspects of anticipation in comprehension, such as how anticipation is modulated by individual difference variables. Comprehenders seem to be able to use words as "cues" to world knowledge, which then enables them to estimate the likelihood of upcoming events (DeLong et al. 2005). Based on recent research, it is now thought to be not intra-lexical spreading as much as event knowledge or schematic knowledge that influences anticipation (Traxler 2014), which is where research on extra-linguistic factors ties in with general comprehension research. If anticipation is indeed based on higher-level knowledge than just that of lexical properties, extra-linguistic information presents an excellent window into those higher levels of anticipation in language comprehension.

Summing up, anticipatory processes seem to be at play in language comprehension. Contextual information appears to be integrated rapidly, and not as a second step after syntactic information is considered. Based on all information available, upcoming items seem to be

[9] It has to be noted that the terminology around anticipatory processing, especially in earlier literature, is far from transparent. Some researches use the terms "anticipation," "facilitation," and "prediction" to clearly distinguish the assumed underlying processes, whereas others use a subset of these term interchangeably, without making a statement on whether the assumed processes are anticipatory, facilitatory, or predictive (Hanulíková et al. 2012; Van Berkum et al. 2005). For those researchers that *do* make a distinction between the terms, *prediction* commonly denotes a process that can on its own generate new candidates for consideration. *Anticipation* and *facilitation*, on the other hand, can adjust activation levels of existing candidates, but cannot generate new candidates by themselves. In this dissertation, I will use the term "anticipation" to subsume anticipatory and facilitatory processes, as distinguished from "prediction", which – by itself – has the power to generate new candidates that were not previously activated for consideration. This discussion is intentionally kept short as the experiments reported below do not have the capacity to distinguish between anticipatory and predictive processes.

anticipated (Federmeier 2007; Kamide et al. 2003; Levy 2008; Sedivy et al. 1999; Tanenhaus et al. 1995; Traxler 2014; Van Berkum et al. 2005; Van Petten and Luka 2012), with anticipation making use of all available clues, such as linguistic and extra-linguistic prior knowledge, event schemas, and real-world information (Altmann and Kamide 1999; Traxler 2014; Van Berkum et al. 2005; however, see Huettig and Mani 2016 and Huettig 2015 for an opposing view, proposing that the experimental effects that suggest anticipation might essentially be task effects, and that anticipation – even if it might occur under certain circumstances – is by no means necessary for language comprehension.)

The four experiments reported below hope to add more findings to the growing body of research, identifying further extra-linguistic variables that may affect language comprehension, by way of linguistic anticipation. Experiment IV specifically investigated the link between Disgust Sensitivity and language comprehension, as one instance of a variable that has been found to affect general cognition, but has not yet been investigated in regards to language comprehension. Even though a detailed theoretical discussion goes beyond the scope of this dissertation, a potential "unifying" theory between anticipation in cognition and in language comprehension could be *Bayesian processing*, which has been found to mesh with empirical results in the fields of cognition (Griffiths et al. 2010), perception, motor control (Knill and Pouget 2004), and neuroscience (Hinton 2007). Very generally speaking, Bayesian processing assumes that the likelihood of a number of interpretations, based on the available cues, is evaluated at the same time (Traxler 2014). In such a scenario, various intra- and extra-linguistic factors could serve as cues to drive likelihood estimation. The assumptions underlying the *Good-Enough* parsing theory, for example, are compatible with a Bayesian approach (Traxler 2014): Under both accounts, interpretations can be biased towards plausible meanings even if syntax technically does *not* allow for these interpretations. That is, it is not just syntactic cues that drive comprehension and anticipation, but also extra-linguistic cues – to a point where those may be more important than syntactic cues. In Bayesian processing, this could simply be explained with a shift in likelihood caused by a number of cues, such as prior discourse or world knowledge, "converging," and hence overpowering the syntax cue (see e.g. Molinaro et al. 2016; Nieuwland and Van Berkum 2006). It should be noted that there are various other general cognition theories available in addition to Bayesian processing; a detailed discussion of the experimental results of this study with regards to various specific cognitive theories is, however, beyond the focus of this dissertation.

1.3 Research Questions – Where to from here?

Based on prior research, this dissertation has as its goal to investigate the influence of various aspects of a listener's identity (namely their personality, political values, and Disgust Sensitivity), combined with aspects inferred by the listener about the speaker's identity (specifically, their gender), on language comprehension. The focus is on the effects of personality traits, beyond empathy, on automated language comprehension – a combination that has been woefully under-researched. In addition, the final experiment will investigate the influence of Disgust Sensitivity, which, to the best of my knowledge, has not yet been researched in regards to linguistic processing.

The study hence addresses the following research questions in particular:

- Do a listener's personality and/or political values influence spoken language comprehension, especially the way they perceive morpho-syntactic errors, semantic anomalies, and socio-cultural clashes?
- Do those extra-linguistic variables influence the automated processing of morpho-syntactic and semantic anomalies differently than they do socio-cultural clashes?
- Does the listener's Disgust Sensitivity, as a marker of BIS activity, modulate language comprehension?

Answers to the above questions, in the form of experimental outcomes, can help to add more experimental evidence to more general questions regarding language comprehension:

- What kinds of extra-linguistic information are considered in language comprehension?
- Specifically, how do listener-internal variables, and information that is inferred about the speaker, modify language interpretation?

To investigate these questions from from multiple angles, and, crucially, to minimize the unintended effects of artifacts or task effects introduced by one particular experimental method, the experiments reported below used a multi-methodological array of studies (see also Arppe and Järvikivi 2007b). This dissertation thus aims at providing more research into which extra-linguistic influences specifically play a role in language comprehension, and how they modify anticipatory processes. Four experiments will be reported below, assessing different aspects of how listener-internal and speaker-inferred variables interact with language comprehension.

The ratings experiment in Chapter 2 investigates off-line language comprehension and allows for an assessment of how extra-linguistic factors influence conscious item ratings. Those item ratings, in addition to being analyzed in their own right, are also included as a predictor in the remaining three experiments. In Chapter 3, a self-paced listening study is used to investigate language comprehension as it happens. As both of these two experiments require a task, and involve conscious movement or decision-making on part of the participants, two pupillometry studies will be reported in Chapters 4 and 5. In the pupillometry paradigm, no task is required, minimizing the influence of potential task effects. In addition, the response variable – pupil size – is not under conscious control of the participants, so that the final two experiments allow a more unmediated look at on-line language comprehension.

Before moving on to the first experiment, Section 1.3.1 below will investigate the participant samples for all four experiments for correlations between extra-linguistic variables (such as personality traits and political values), and compare the distributions of Big Five traits (John and Srivastava 1999; John et al. 1991; also see Section 2.4 and Table 2.2) to those reported in prior literature.

1.3.1 Correlations between & Distributions of Personality Traits

While some weak to moderate correlations between certain variables were found, there is no consistent pattern (cf. Table 1.2): Only three pairs of values were found to correlate with a coefficient higher than |.3| more than once, namely Openness and Extraversion (.37 and .32 in Exp. II, III, and IV), Conscientiousness and Agreeableness (.35, .36, and .33 in Exp. I, III, and IV), and Agreeableness and Neuroticism (-.39 and -.32 in Exp. I and III). Interestingly, the participants' political values score was only found to correlate moderately with Openness and Conscientiousness in one of the four experiments (.28 in Exp. III). This overall pattern of rather low correlations is in line with findings presented in previous literature: For example, Table A1 in Gerber et al. (2010) notes that all correlations between

Big Five traits were found to be under .40 in value, with the two strongest correlations observed between Conscientiousness and Stability (the inverse of the Neuroticism scale), and Agreeableness and Stability.

Some prior literature suggests that Openness is correlated with a progressive leaning, and Conscientiousness with a conservative approach (Gerber et al. 2010; Webster 2018). Cawvey et al. (2016) have identified that the link between Openness and Liberalism is among the two most commonly observed correlations between personality traits and political values, together with the link between Conscientiousness and Conservatism. In the four experiments in this dissertation, only weak-to-moderate correlations were found between Openness and progressiveness in the second and third experiments (.20 and .28, respectively; again refer to Table 1.2), whereas Conscientiousness, in contrast to findings from prior literature, *also* trended towards a more progressive leaning – and only in one of the three experiments (Exp. III, .28). Neuroticism was not consistently associated with a correlation in either political direction. In summary, no consistent correlations between political values and any of the Big Five traits were identified across the four experiments in this dissertation. As per Verhulst et al. (2013) and Hatemi and Verhulst (2015), where changes in personality across time did not predict changes in political attitudes, any correlations should not be assumed to be the result of a causal relationship in which personality traits "trigger" a particular political outlook, but rather the result of a common underlying causal variable, such as genetic predisposition. Of note is that the highest correlation observed across all four experiments was .40, i.e. a moderate positive correlation, between Agreeableness and Disgust Sensitivity in Experiment IV, a correlation that, as discussed in Section 1.1.2.4, has been attested in the literature (Druschel and Sherman 1999).

To ensure that the samples of university students were not significantly different from trait distributions in the general population, the distributions of Big Five scores in all four participant samples was compared to several other Big Five distributions reported in the literature (see Table 1.1 and Fig. 1.1). The student samples from all four experiments were found to be well in line with historical distributions, and were found to be closest to the two largest data sets in the literature: Firstly the ISDP data, from a large-scale survey on sexuality with data from more than 40 countries in 24 languages (Schmitt and Shackelford 2008), and secondly the data in Srivastava et al. (2003), which was sampled from a large number of North American adults of varied ages. Gurven et al. (2013) was intentionally included as a bit of an "outlier," as the data was collected from indigenous Bolivian forager-farmers, a distinctly non-urbanized, non-Western population. Rammstedt (2007) analyzed a German adult population.[10]

We will now move on to the first of four experiments, which investigated the effect of listener- and speaker-related variables on conscious, off-line item ratings.

[10] Note that "Neuroticism" is referred to as "Emotional stability," its inverse, in this data set.

		Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
	$n =$	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Boland and Queen (2016)	83	3.68	0.64	3.92	0.71	3.05	0.77	3.69	0.62	2.48	0.87
Gurven et al. (2013)	632	3.01	0.46	3.03	0.5	2.53	0.48	3.41	0.44	2.44	0.39
ISDP (Schmitt and Shackelford 2008)											
men	5,445	3.71	0.6	3.39	0.66	3.32	0.68	3.57	0.59	2.79	0.72
women	7,798	3.68	0.6	3.5	0.66	3.43	0.73	3.68	0.6	3.14	0.75
Rammstedt (2007)	2,569	3.41	0.88	4.1	0.69	3.24	0.88	3.2	0.83	3.49	0.85
Srivastava et al. (2003)											
age: 21	6,076	3.92	0.66	3.45	0.73	3.25	0.9	3.64	0.72	3.23	0.82
age: 25	3,683	3.96	0.66	3.58	0.71	3.31	0.91	3.66	0.71	3.27	0.83
This Dissertation											
Item ratings	125	3.58	0.58	3.37	0.59	3.15	0.81	3.92	0.53	3.13	0.78
Self-paced listening	83	3.49	0.63	3.05	0.59	2.44	0.73	3.32	0.57	2.67	0.63
Pupillometry	39	3.37	0.50	3.33	0.59	3.04	0.76	3.78	0.45	3.20	0.65
Pupillometry & Disgust	82	3.55	0.70	3.58	0.59	3.15	0.86	3.86	0.58	3.25	0.73

Table 1.1: A comparison between Big Five distributions in the literature and those observed in the four experiments of this study. Cross-reference with Fig. 1.1.

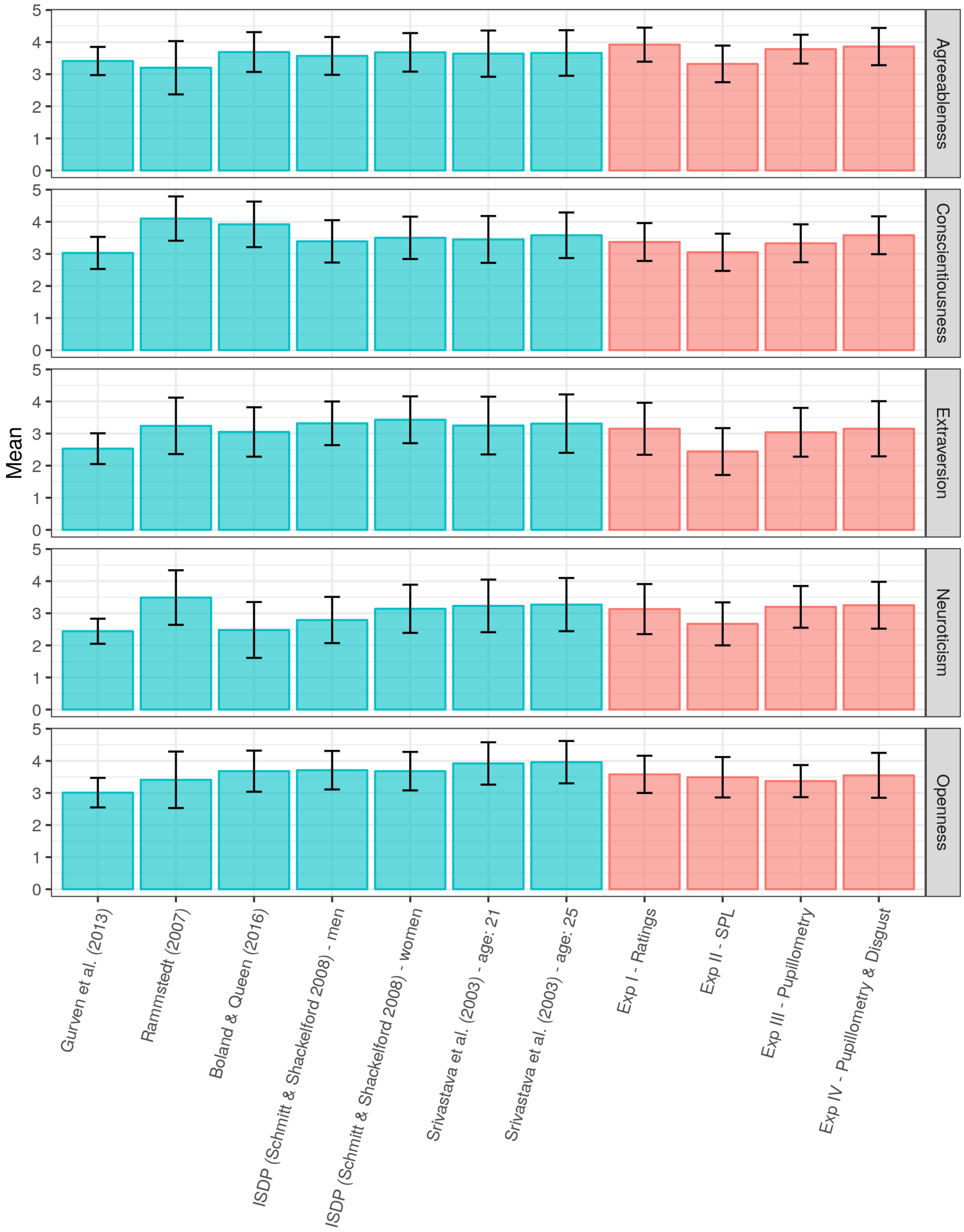


Figure 1.1: A visual comparison of Big Five distributions in the literature and those observed in the four experiments of this study. Error bars denote 1 SD from the mean. Cross-reference with Table 1.1.

		Consc.	Extr.	Agr.	Neur.	Pol.	Disgust
Exp. I	*n = 99*						
	Openness	-.10	.24	.17	0	.11	
	Conscientiousness		.19	.35	-.31	.04	
	Extraversion			.19	-.29	.02	
	Agreeableness				-.32	.15	
	Neuroticism					.06	
Exp. II	*n = 43*						
	Openness	.22	.37	.20	-.08	.20	
	Conscientiousness		.16	.11	-.14	.02	
	Extraversion			.26	-.13	-.04	
	Agreeableness				-.29	.23	
	Neuroticism					-.02	
Exp. III	*n = 48*						
	Openness	.34	.32	.17	.08	.28	
	Conscientiousness		.12	.36	-.13	.28	
	Extraversion			.03	.14	-.04	
	Agreeableness				-.39	.17	
	Neuroticism					.14	
Exp. IV	*n = 76*						
	Openness	.09	.37	-.04	-.28	-.10	-.09
	Conscientiousness		.15	.10	.08	-.05	.13
	Extraversion			.33	-.15	-.06	.20
	Agreeableness				-.12	.11	.40
	Neuroticism					-.12	.29
	Political values						.23

Table 1.2: Correlations observed between Big Five personality traits, political values, and Disgust Sensitivity (where applicable) across Experiments I through IV. Correlation coefficients larger than a value of .3 are highlighted in orange, with coefficients larger than a value of .25 highlighted in yellow. Note that the political values scale in Experiment IV uses the opposite polarity compared to the scale used in Experiments I through III. That is, in Experiment IV, high political values scores indicate a conservative leaning.

Chapter 2

Experiment I: Item Ratings

In this first experimental chapter, we will investigate whether a listener's personality or gender, or the speaker's (inferred) gender, influences the acceptability ratings of anomalous utterances as compared to a non-anomalous baseline. In the main experiment, participants were presented with the stimuli in auditory format (for details, see Section 2.2 below), and rated them for acceptability on a four-point ratings scale. As ratings were gathered after the participant heard each item, in a non-time-locked fashion, this experiment is comparatively more off-line than the other three experiments in this dissertation, and presents a look at the more conscious aspects of language comprehension. In addition to being analyzed as an experiment in its own right, this ratings study also provided the average item ratings that were used as a numerical predictor in the analysis of the other three experiments in this dissertation. Numerical ratings were preferred as a statistical predictor over a simple binary clashing/non-clashing distinction, as they provide a more fine-grained assessment in the form of a numerical scale. In addition to semantic anomalies (such as "Dogs often chase *teas*...") and morpho-syntactic errors (such as "He often *walk* his dog..."), the focus in this study (as in the whole dissertation) is also on the listener's response to socio-cultural clashes – statements clashing with gender stereotypes as inferred from the male or female voice of the Canadian English speaker, such as "I buy my *bras*..." spoken by a male speaker (for details on stimuli and the three different clash types, see Section 2.2 below). We predicted that socio-cultural clashes, as per the nature of this type of clash, were likely going to show more variance in ratings compared to semantic violations and morpho-syntactic errors, precisely because we anticipated that ratings were going to be influenced by the raters' real-world experience and personality. After the main experiment, the participants' personality was assessed using the Big Five personality assessment (John and Srivastava 1999; John et al. 1991; for details, see Table 2.2 and Section 2.4).

2.1 Participants

125 students, recruited from the pool of undergraduate linguistics students at the University of Alberta, participated in this experiment. Non-native speakers of English were awarded credit for their participation, but their data was not used in the analyses. As a result, the analyses below are based on the data obtained from 99 native speakers of North American English (males/females = 59/40; age = 17–31; mean = 20.4 years).

2.2 Materials

			Segment		
	I	II	**III - Critical**	**IV - Post-Critical**	**V - Wrap-Up**
MO	She	usually	drives/**drive** her car	slowly	in the snow.
SE	People	often	read books/**heads**	for pleasure	at night.
SC	I	always	enjoy *knitting/football*	in my	free time.

Table 2.1: The template used for item construction, with three example sentences. MO = morpho-syntactic errors; SE = semantic anomalies; SC = socio-cultural clashes.

240 sentence stimuli were created, distributed among the following conditions (examples are given in Table 2.1):

Morpho-syntactic errors: 56 stimuli in total, half of which violated subject-verb agreement (De Vincenzi et al. 2003; Ditman et al. 2007);

Semantic anomalies: 32 stimuli in total, half of which contained a semantic mismatch between the verb and the object (De Vincenzi et al. 2003; Ditman et al. 2007);

Socio-cultural clashes: 120 stimuli in total, half of which contained a clash with the speaker's perceived identity as per common gender stereotypes (as per e.g. Van Berkum et al. 2008; Van den Brink et al. 2010);

Unrelated fillers: 32 non-anomalous filler sentences, such as “Chickens normally live in a coop.”

The sentences all followed the same syntactic pattern to ensure comparability across regions (Jegerski and VanPatten 2013; again see Table 2.1 for the template, and Appendix A for the full list of items). For item recording, items were presented to one male and one female native speaker of Western Canadian English in random order. In line with practices reported in Van Berkum et al. (2008), item recordings in which the prosody sounded noticeably different from those of other items were re-recorded with the speaker.[1] Items were then distributed across four lists of 135 items each, counterbalanced for error condition (correct/non-anomalous vs. anomalous/clashing) and speaker (male vs. female), and each participant was presented with one list (and, accordingly, each item only once, in just one condition and spoken by one speaker.)

2.3 Procedure

After a short briefing, participants were seated at a desktop computer and asked to wear the headphones provided. They were then presented with one out of the four lists of items, to avoid repeated exposure to the same stimulus spoken by different speakers. Participants were then given the chance to ask questions after a short practice section. During the experiment, each stimulus was played to the participant, accompanied by “How does this statement sound to you?” printed on the screen. Participants were asked to rate the acceptability of the stimulus via mouse-click on a four-point ratings scale, from zero (“not acceptable”) to three (“fully acceptable”). Participants were instructed to interpret “acceptability” to

[1] While recording both non-anomalous and anomalous items directly with a speaker may cause acoustic properties in their speech to hint at an upcoming problem, recording stimuli in this fashion is common practice in current individual differences research. Additionally, if differences in prosodic contours or other acoustic properties indeed hint at an upcoming clash, we would also expect those hints to be a component of natural speech in the real world, outside of the laboratory; Conversely, alternatives such as splicing and synthetic speech generation may not hint at an upcoming “problem,” but could introduce other strange features entirely unrelated to the issue under investigation, thus causing a different set of issues. Additionally, if prosody hints at an upcoming clash, we may expect either significant increases in pupil size *before* the actual clash, or a non-significant change in pupil size after the clash, as listeners may already have been expecting it as per the incoming acoustic signal. As will be shown in Chapters 4 and 5, results do not support either of these notions.

refer not just to ungrammatical utterances, but to any aspect of the utterance that makes it sound “strange” to them; however, still note the discussion in Penke and Rosenbach (2004) regarding differences in acceptability judgments.

Different Likert-style or ratings point scales have been in use in recent research, ranging from e.g. five (Haapalainen et al. 2010) to seven (Boland and Queen 2016; Grey and Van Hell 2017), nine (Ahn et al. 2014), or even eleven (Molinaro et al. 2016) points. A smaller scale was chosen for this ratings experiment as participants were not asked about personal agreement with, for example, a political statement, or to judge a person’s character, which would necessitate a more fine-grained scale; rather participants were asked to assess sentences along a simpler dimension, namely their acceptability.

2.4 Post-Tests

Sub-scale	Traits associated with high scores
Agreeableness	cooperative, trustful, sympathetic
Conscientiousness	orderly, responsible, dependable
Extraversion	talkative, assertive, energetic
Neuroticism	easily upset, neurotic, not calm
Openness	curious, creative, unconventional

Table 2.2: An overview of the Big Five sub-scales (John and Srivastava 1999; John et al. 1991) used to assess the participants’ personality, and traits associated with high scores on the respective scales.

After the main ratings experiment, so as to not prime participants towards the purpose of the study, personality traits were assessed via the *Big Five* (John and Srivastava 1999; John et al. 1991) personality inventory, coded in `E-Prime` (Psychology Software Tools Inc. 2012). An overview of the test and its subscales, with examples of associated traits, is provided in Table 2.2, and the full test can be found in Appendix B.2. The Big Five inventory was chosen for its frequent and continued use in psychological research, and/or because it assesses various aspects of an individual’s personality rather than just providing one overall score. The participants’ political values were assessed via a *Political Ideology Questionnaire*, created by the School of Social Work at Louisiana State University (Grenier n.d. which also formed the basis for the assessment of political values in Marrville 2017), again coded in `E-Prime` (Psychology Software Tools Inc. 2012). The full test can be found in Appendix B.3.

High scores on this test indicate a progressive stance, and lower scores are associated with a more conservative outlook. Data on the participants' language background was collected via a pen-and-paper language background questionnaire, which can be found in Appendix B.5.

2.5 Results

All results reported below were obtained through linear mixed effects regression modelling (LMER) in `R` (R Core Team 2019, version 3.5.3), using the `lme4` (Bates et al. 2015, version 1.1-21) and `lmerTest` packages (Kuznetsova et al. 2017, version 3.1-0). For further analysis, reporting, and visualization, the `effects` (Fox and Hong 2009, version 4.1-0), `stargazer` (Hlavac 2018, version 5.2.2), `ggplot2` (Wickham 2016, version 3.1.0), and `MuMIn` (Bartoń 2018, version 1.42.1) packages were used. The dependent variable in all models is item rating, and each model includes random intercepts for participant and item. Changes in ratings over time, that is, across the experiment, were tested (cf. Divjak et al. 2016), but not found to be significant in any of the models reported below. Models were fitted and selected using a backwards step-wise elimination procedure, comparing each iteration and testing for significance of the main effect or interaction in question via ANOVAs and the Akaike Information Criterion (AIC). Random by-item-and-participant intercepts and slopes, as well as random by-participant-and-personality-trait intercepts and slopes, were tested, however there were not enough observations in the data to support this kind of modelling structure.

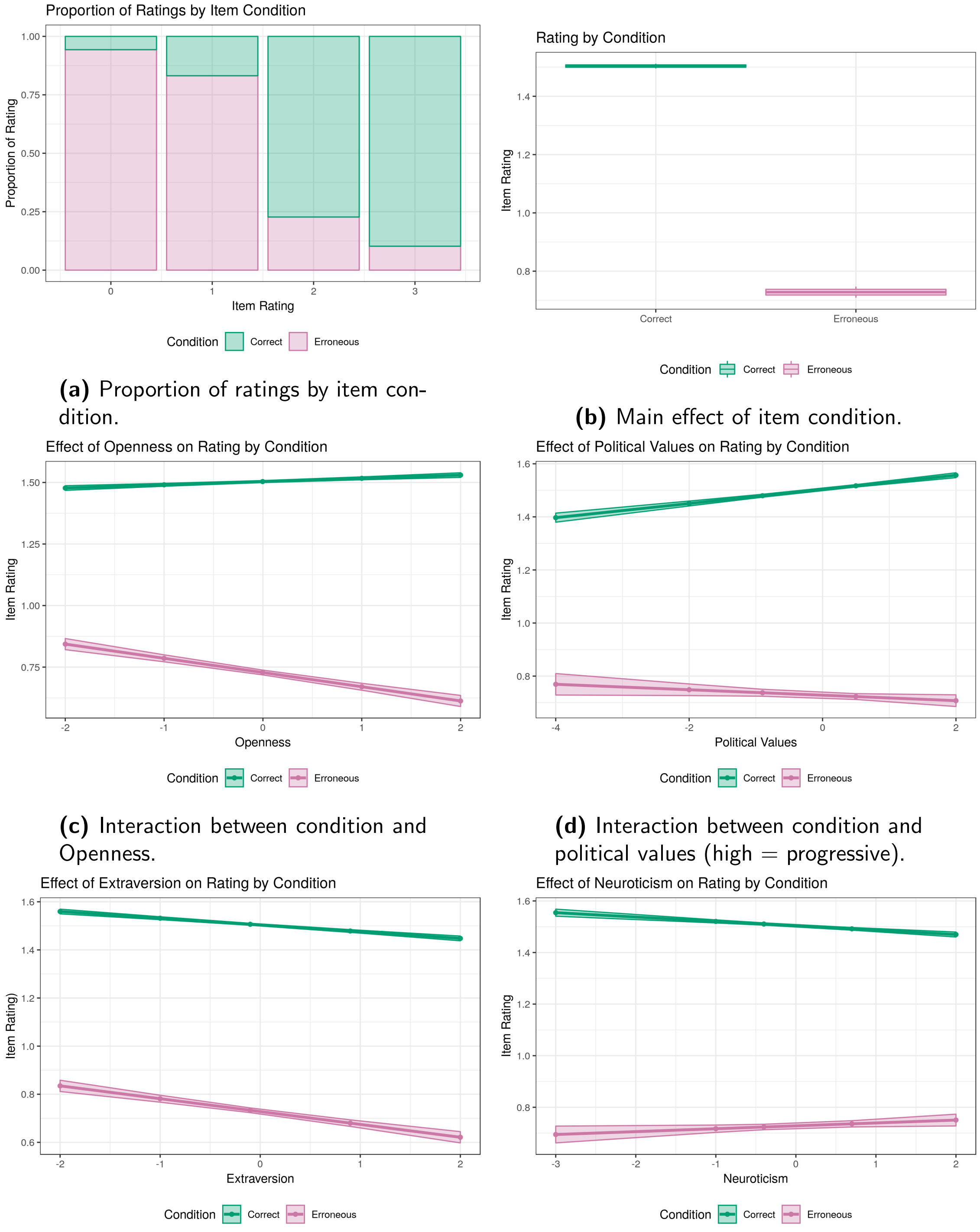
2.5.1 Morpho-Syntactic Errors

For the morpho-syntactic error type, the best model fit ($AIC = 6802.647$, marginal $R^2 = .37$, conditional $R^2 = .48$; for the full model output, see the leftmost column in Table 2.3) found a **main effect of condition**, with erroneous items being rated significantly lower on the acceptability scale than correct items, as expected (cf. Figs. 2.1a and 2.1b).

Of particular interest for this dissertation are the four significant two-way interactions that were found, namely between: condition and Neuroticism; condition and Extraversion; condition and political values; and condition and Openness. We will discuss each of these interactions in turn below.

	Effects on Rating		
	Morpho-syntactic	**Semantic**	**Socio-cultural**
Number of observations	$n = 9,482$	$n = 8,966$	$n = 11,098$
(Intercept)	1.504 $p < 0.001^{***}$	1.495 $p < 0.001^{***}$	1.497 $p < 0.001^{***}$
Condition	−.776 $p < 0.001^{***}$	−.683 $p < 0.001^{***}$	−.111 $p < 0.001^{***}$
political values		.028 $p = .024^{*}$	
Speaker gender			.015 $p = .014^{*}$
Condition : Neuroticism	.028 $p = .007^{**}$		
Condition : Extraversion	−.025 $p = .016^{*}$	.042 $p = .0004^{***}$	
Condition : Political values	−.037 $p = .0002^{***}$	−.121 $p = 0.000^{***}$	−.025 $p = .0004^{***}$
Condition : Openness	−.071 $p = 0.000^{***}$		.019 $p = .007^{**}$
Condition : Agreeableness		.056 $p = .00002^{***}$	
Condition : Speaker gender		−.062 $p = .013^{*}$	
Condition : Listener gender		.168 $p = 0.000^{***}$	
Listener gender : Neuroticism		−.063 $p = .016^{*}$	
Openness : Speaker gender			.014 $p = .021^{*}$

Table 2.3: LMER output for the three clash types in Experiment I. Each row shows *estimates* for each predictor, with the *significance level / p-value* just below. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Note that ratings were square root transformed for modelling, and all numerical predictors were scaled and centered. Only predictors significant at the .03 level are shown.



(a) Proportion of ratings by item condition.

(b) Main effect of item condition.

(c) Interaction between condition and Openness.

(d) Interaction between condition and political values (high = progressive).

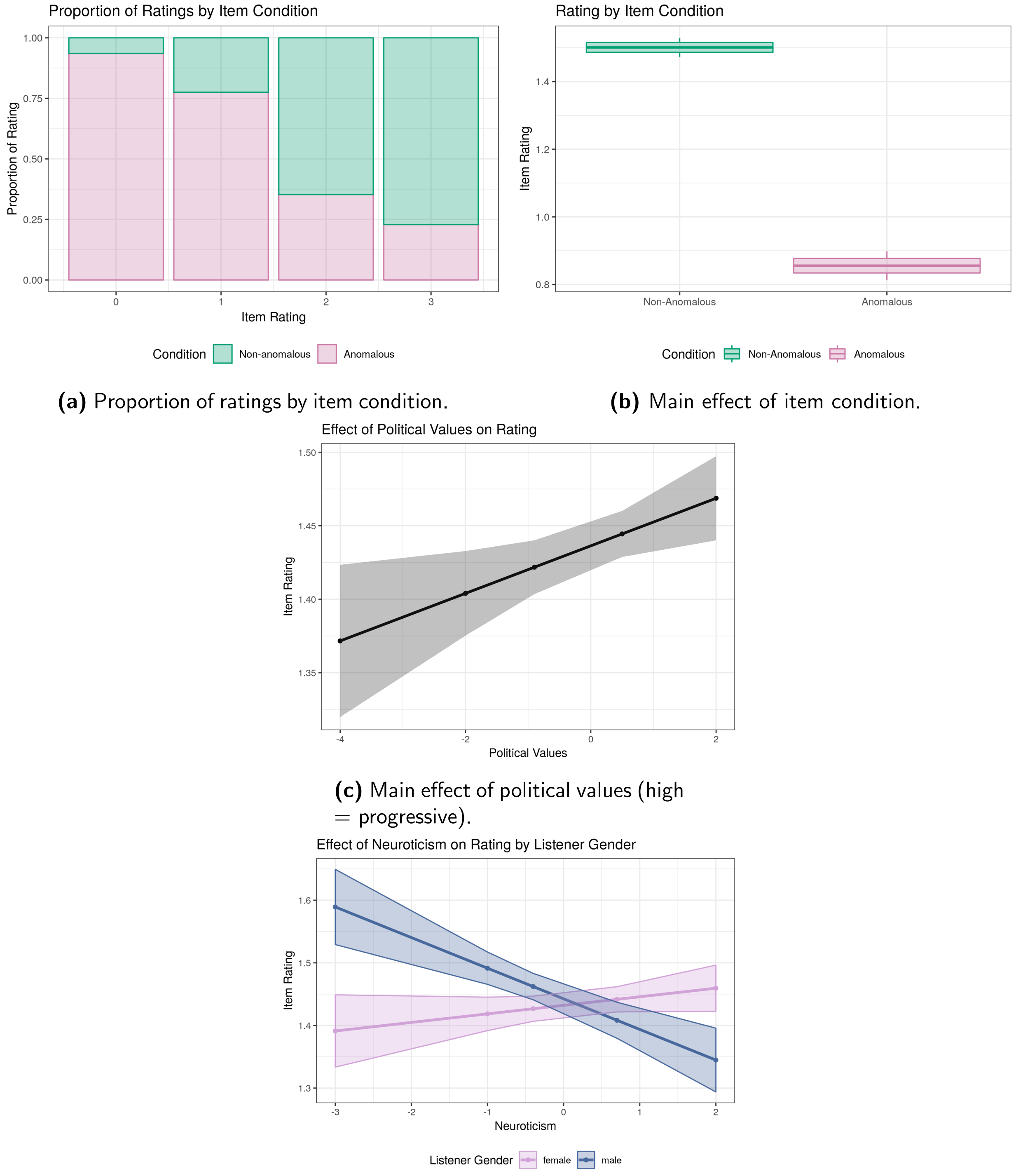
(e) Interaction between condition and Extraversion.

(f) Interaction between condition and Neuroticism.

Figure 2.1: Visualization of ratings distribution and effects in the LMER model for morpho-syntactic errors. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Ribbons in line plots indicate the default standard error as calculated via the `effects` package; additional vertical lines in boxplots denote the 95% confidence interval. Higher item ratings mean higher acceptability. Note the differing y-axes.

In two of those interactions, namely between **condition and Openness**, and **condition and political values**, which are visualized in Figs. 2.1c and 2.1d, i.e. in the middle row of Fig. 2.1, individuals with higher Openness and political value scores (on the x-axis in the visualizations) rated correct items better, and erroneous items worse than individuals with lower scores on those scales; compare the red and green slopes in the visualizations. Thus, more open and more liberal/progressive individuals exhibited a similar pattern in their item ratings. Given that highly open individuals are described as curious, creative, and unconventional (cf. 2.2), and could be expected to not be "thrown off" as much by morpho-syntactic errors, this pattern is not intuitively accessible. It is the opposite of what Boland and Queen (2016) have found in response to written errors, although it has to be noted that in their study, participants were asked to rate the prospective housemate (i.e. the author of the text that contained the error), which is not immediately comparable to rating the acceptability of an error itself.

In the remaining two interactions, between **condition and Extraversion**, and **condition and Neuroticism**, higher scores on the respective scale meant lower ratings for correct items, as visualized in Figs. 2.1e and 2.1f, i.e. in the bottom row of Fig. 2.1: Highly extraverted and neurotic listeners rated correct items worse than introverted and less neurotic individuals. However, the two interactions differ in the ratings pattern for erroneous items: **extraverted** individuals rated erroneous sentences worse than introverts did; this effect was stronger for erroneous sentences as compared to correct ones (cf. the steeper slope for erroneous sentences in Fig. 2.1e). Highly **neurotic** individuals rated erroneous items *better* than their less neurotic peers (cf. Fig. 2.1f); as such, less neurotic listeners made a bigger difference in ratings between correct and erroneous sentences than their more neurotic counterparts, for whom ratings for correct and erroneous items were closer to each other. This suggests that more neurotic individuals may be less certain of their judgment, which may be a reflection of more neurotic individuals generally being less self-confident (see e.g. John and Srivastava 1999).



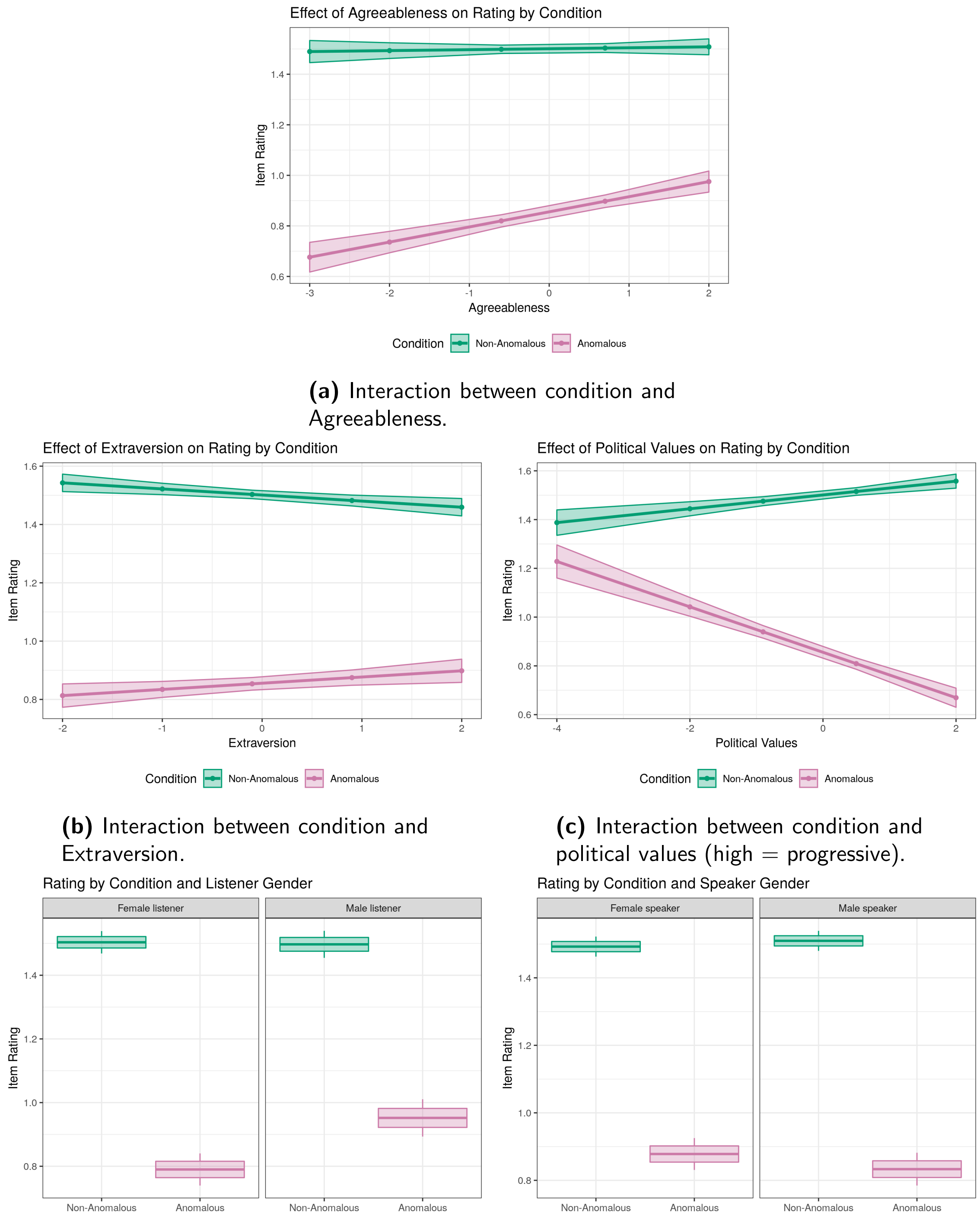
(a) Proportion of ratings by item condition.

(b) Main effect of item condition.

(c) Main effect of political values (high = progressive).

(d) Interaction between Neuroticism and listener gender.

Figure 2.2: Visualization of ratings distribution and effects in the LMER model for semantic anomalies; continued in Fig. 2.3 below. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Ribbons in line plots indicate the default standard error as calculated via the `effects` package; additional vertical lines in boxplots denote the 95% confidence interval. Higher item ratings mean higher acceptability. Note the different y-axes.



(a) Interaction between condition and Agreeableness.

(b) Interaction between condition and Extraversion.

(c) Interaction between condition and political values (high = progressive).

(d) Interaction between condition and listener gender.

(e) Interaction between condition and speaker gender.

Figure 2.3: Visualization of further effects in the LMER model for semantic anomalies; continuation of Fig. 2.2 above. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Ribbons in line plots indicate the default standard error as calculated via the `effects` package; additional vertical lines in boxplots denote the 95% confidence interval. Higher item ratings mean higher acceptability. Note the different y-axes.

2.5.2 Semantic Anomalies

For semantic anomalies,[2] just as for the morpho-syntactic errors discussed previously, the best model (AIC = 6333.103, marginal R^2 = .24, conditional R^2 = .36; for the full model output, see the middle column of Table 2.3) again found a **main effect of item condition**, with erroneous items being rated significantly lower for acceptability (cf. Figs. 2.2b and 2.2a). The model also found a **main effect of political values**, where more liberal/progressive individuals generally rated all items better overall, irrespective of condition (cf. Fig. 2.2c). Item condition interacted significantly with several extra-linguistic variables, such as Extraversion, Agreeableness, political values, speaker gender, and listener gender (see all panels in Fig. 2.3). Unlike for morpho-syntactic errors above, no highly similar ratings patterns were observed between the two-way interactions of condition and an extra-linguistic variable in the model for semantic anomalies. We will now discuss each of these two-way interactions in turn, beginning with those interactions that involve item condition.

More **agreeable** individuals were found to rate anomalous sentences better than their less agreeable peers (cf. the red slope in Fig. 2.3a); at the same time, agreeableness score did not seem to influence ratings for non-anomalous sentences significantly (cf. the green slope in the same figure). Highly agreeable individuals thus showed less of a discrepancy in ratings for correct and anomalous sentences, as shown by the smaller gap towards the right side of Fig. 2.3a.

In the interaction between **condition and Extraversion**, highly extraverted individuals rated anomalous sentences higher than their introverted peers (cf. the red slope in Fig. 2.3b), similarly to the effect of Agreeableness just discussed. However, Extraversion seemed to affect the ratings of *non-anomalous* sentences as well (cf. the green slope in the same figure): Extraverted listeners rated correct sentences *worse* than their introverted counterparts, thus making less of a difference in item ratings between the two conditions than introverted listeners. Note how, compared to this same interaction (between condition and Extraversion) in the model for morpho-syntactic errors, the slope for erroneous items runs in the *opposite* direction here (compare Figs. 2.1e and 2.3b). This suggests that the processing of morpho-syntactic errors and semantic anomalies is influenced by personality in different ways: Whereas for morpho-syntactic errors, a listener's higher Extraversion score meant

[2] Recall that semantic anomalies were mismatches between the verb and its object, such as "Dogs sometimes chase *teas* on the road for fun" (see also Table 2.1).

lower ratings for *both* correct and erroneous items, a higher Extraversion score meant higher ratings for semantically anomalous items. It is conceivable that, while morpho-syntactic errors always remain errors (at least if the speaker sounds like a native speaker, as was the case in this experiment – see also Hanulíková et al. 2012), that semantic errors "lose some of their edge" when listeners are exposed more to unusual stimuli. This increased exposure to unusual stimuli can be assumed for highly extraverted listeners, who generally seek out social interactions more than introverts, and would hence be exposed more to unusual linguistic stimuli. Another explanation could be that extraverted individuals may make more use of cues they were able to derive about the speaker, simply because they can be assumed to be more adept at using social cues due to more frequent social encounters. They might hence engage in more anticipation based on the speaker's native accent in this case, which would go hand in hand with an absence of morpho-syntactic errors. As such, highly extraverted listeners might then experience more surprisal when the speaker does indeed produce a morpho-syntactic error, and hence rate the item as less acceptable. Semantic anomalies, on the other hand, do not lend themselves to this kind of extrapolation - dogs hunting teas instead of toys or cats can hardly be anticipated based on any inferred characteristic of the speaker. As such, they may trigger less surprisal, resulting in better ratings. Note that this ratings experiment cannot distinguish between these two hypotheses; findings will be discussed in context of the other three experiments in the General Discussion (Chapter 6).

In the interaction between **item condition and political values**, more liberal/progressive individuals, i.e. those with higher scores on the political values scale, rated correct items better, and anomalous items much worse than their more conservative peers (cf. Fig. 2.3c). This is in line with the findings from the morpho-syntactic model discussed above, which found a highly similar effect pattern, albeit less pronounced (refer back to Fig. 2.1d). These findings suggests that liberal individuals have a wider "scale" on which they rate items compared to conservative individuals (compare the size of the gap between the slopes for the two different conditions in both figures); conservative listeners seem to be more restrained in their ratings, not opting for the extreme ends of the ratings scale.

In the interaction between **condition and listener gender**, male listeners rated anomalous items better than female listeners (cf. the red boxes in Fig. 2.3d), whereas there was no significant difference in how listeners of either gender rated non-anomalous items (cf. the green boxes in the same figure). Listeners of either gender, as expected, rated anomalous sentences much worse than non-anomalous items.

In the interaction between **condition and speaker gender,** there was a tendency for a non-anomalous item to be rated better when it was spoken by a male speaker (cf. the green boxes in Fig. 2.3e). A significant difference was found for anomalous items, which were rated worse when they were produced by a male speaker (cf. the red boxes in the same figure). Interestingly, semantic anomalies are the only clash type for which interactions between item condition and speaker gender, and item condition and listener gender were found.

The effects observed in these two interactions with gender could have to do with the "male as default" setting, the generally larger privileges and prestige of men as compared to women, and the greater attention levels given to male speakers (Gruber and Gaebelein 1979; Orlob 2017). Under these assumptions, it is highly conceivable that correct items would be rated worse when they were produced by a female speaker as compared to a male speaker, and that *erroneous* items would be rated *worse* when the speaker was male. As per the assumption of male as default, a male speaker may not be expected to produce something "irrational" and "unpredictable" as a semantic anomaly. Listeners heard both speakers produce the same percentages of correct and erroneous items, so it is unlikely that the effect would have its origin in surprisal stemming from the likelihood of a speaker uttering an anomalous sentence.

Finally, a significant interaction was observed between the listener's **Neuroticism score and their gender** (note that item condition is not a factor in this interaction): Whereas more neurotic female listeners showed a tendency to rate items better than their less neurotic peers (cf. the pink-ish slope in Fig. 2.2d), male listeners rated items significantly *worse* the more neurotic they were (cf. the blue slope in the same figure). Two interesting findings emerge from this: Firstly, Neuroticism scores seemed to influence ratings much more among male listeners; and secondly, items were rated worse by less neurotic listeners when the listener was female, but rated worse by *highly* neurotic listeners when the listener was male (note the slopes crossing over each other in Fig. 2.2d). An interpretation of this crossover effect is not immediately available, and more research is needed to investigate the interplay between Neuroticism and listener gender with regards to language comprehension. However, it should be noted that women on average are more neurotic than men (Ormel et al. 2013), so that highly neurotic men can be considered stronger outliers than highly neurotic women, which may result in unexpected effects. However, in the participant sample for this present study, only a tendency was found for women to be more neurotic than men ($mean_{male} = 3.00, SD_{male} = 0.89; mean_{female} = 3.29, SD_{female} = 0.78; t(80.953) = -1.7331, p = 0.09$).

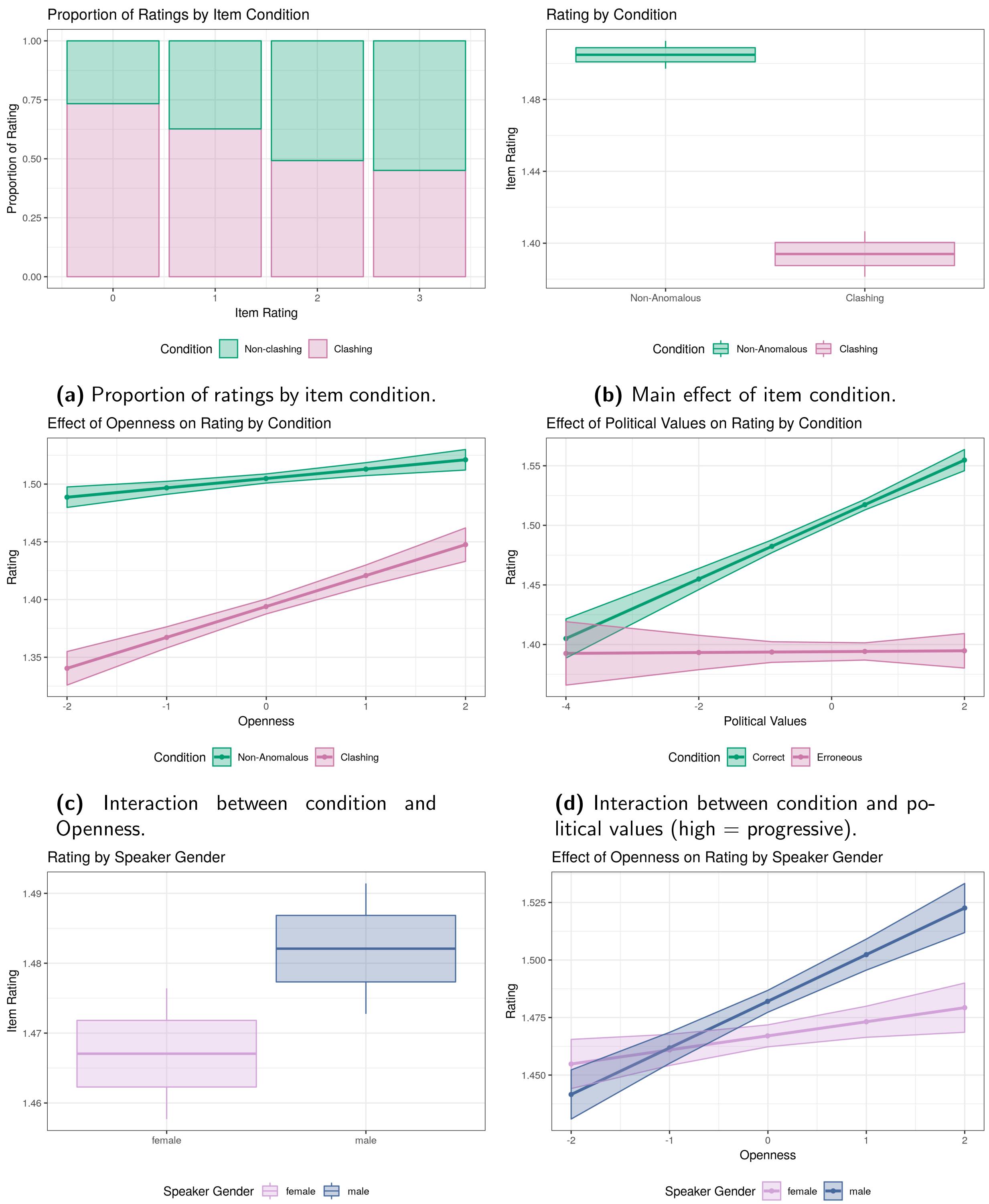
2.5.3 Socio-Cultural Clashes

For socio-cultural clashes, the best model ($AIC = 6648.677$, marginal $R^2 = .03$, conditional $R^2 = .23$; for the full model output, see the rightmost column in Table 2.3) also found a **main effect of condition**, albeit one that was quite a bit smaller than for the other two anomaly types discussed above (cf. Fig. 2.4b and 2.4a). This was expected, given the fact that socio-cultural clashes, as per their nature of clashing with established stereotypes, can be assumed to be influenced more by the real-world experience of a listener than an intra-linguistic error.

Further, a **main effect of speaker gender** was found, where items spoken by a male speaker were generally rated better than those spoken by a female speaker (cf. Fig. 2.4e). This can likely be attributed to the "male as default" pattern that has already been discussed for semantic anomalies in the previous section, where the male gender is generally unmarked, and male speakers are given more attention than female speakers overall (Gruber and Gaebelein 1979; Orlob 2017).

The model found several significant interactions, namely between condition and Openness, condition and political values, and between Openness and speaker gender; we will again discuss all of these in turn, starting with those interactions that involve item condition, as those are of special interest for this dissertation.

In the significant interaction between **item condition and Openness**, more open individuals rated both non-clashing and clashing items better than their less open peers (cf. Fig. 2.4c). Note how the slope (in this same figure) for clashing items is steeper than for non-clashing items; that is, more open listeners rated non-clashing items a little better than their less open peers, but clashing items a *lot* better, comparatively. This effect is as expected – more open individuals, as per their traits, were expected to rate clashes based on



(a) Proportion of ratings by item condition.

(b) Main effect of item condition.

(c) Interaction between condition and Openness.

(d) Interaction between condition and political values (high = progressive).

(e) Main effect of speaker gender on item ratings.

(f) Interaction between Openness and speaker gender.

Figure 2.4: Visualization of effects in the LMER model for socio-cultural clashes. Ribbons in line plots indicate the default standard error as calculated via the `effects` package; additional vertical lines in box plots denote the 95% confidence interval. Higher item ratings mean higher acceptability. Note the different y-axes.

stereotypes as more acceptable than their less open peers. We will return to this interaction, and to how it compares to the same interaction for morpho-syntactic errors, in the discussion below.

In the significant interaction between **political values and item condition**, individuals with lower political value scores (i.e. those leaning towards the conservative side) rated *non-clashing* items much worse than their more liberal counterparts (see the green slope in Fig. 2.4d); individuals at both ends of the scale rated clashing items about the same (see the red slope in the same figure). This shows that extremely conservative listeners made no difference, or only a very minor one, in ratings between clashing and non-clashing items. This presents an "exaggerated" version of the finding in the morpho-syntactic and semantic conditions for this same interaction (refer back to Figs. 2.1d and 2.3c), namely that progressive individuals appear to have a wider "scale" on which to rate items. This does not necessarily mean that progressive individuals are more "judgmental" when it comes to utterances clashing with established gender stereotypes, but rather reflects the much better ratings they awarded to non-anomalous utterances (refer back to the steep, green slope in Fig. 2.4d). At this time, there is no adequate explanation for why conservative individuals would rate non-clashing items so badly that they end up being rated virtually the the same as clashing items; further research on how conservatives navigate the (linguistic) world is needed.

In the final interaction, between **Openness and speaker gender** (note that this interaction is not concerned with item condition), less open listeners rated items spoken by the male and female speaker about the same (cf. the left hand side of Fig. 2.4e). The more open a listener, the better the ratings, as is shown by the two slopes inching upward toward the right side of the same figure. However, the slope for the male speaker was much steeper, suggesting that more open listeners rated items produced by the male speaker *better* than those produced by the female speaker (cf. the gap between the two slopes becoming wider towards the right side of the figure). While this could likely be related to the "male as default" effect already discussed previously (Gruber and Gaebelein 1979; Orlob 2017), it is unclear why this effect would affect more open individuals more than their less open peers. A potential avenue for explanation could be that, based on the speaker's gender as inferred from their voice, more open listeners may engage in more anticipation regarding the "quality" of the utterance: It is possible that a male voice would prime listeners for a better rating, as male speakers are considered the default. Less open individuals may use this social cue stemming from the speaker's voice a little less, and focus more on the linguistic content of the utterance than extra-linguistic cues, and thus rate items equally, irrespective of who produced them. This is a rather non-intuitive and speculative explanation for the effect found here; more research is needed at the intersection of speaker gender and Big Five traits.

2.6 Discussion

Summing up the findings of the ratings experiment, a **main effect of item condition** was found for all three clash types. As expected, the effect was strongest for morphosyntactic errors (Fig. 2.1b), followed by semantic anomalies (Fig. 2.2b), and was weakest – but still significant – for socio-cultural clashes (Fig. 2.4b), mirroring the differences in ratings distributions (cf. Figs. 2.1a, 2.2a, and 2.4a).

In all three clash types, several **interactions between item condition and variables relating to the speaker's or listener's identity** were found: The most pervasive interaction was that with a listener's **political values**, which was significant in all three clash types. This is an interesting observation, as political leaning – while correlated with some personality traits, as discussed in Section 1.3.1 – is influenced more by learned behaviours, societal constructs, and conscious thought than personality traits are. As we will see in Chapter 3, political values were much less pervasive as a predictor across the three clash types in the three other experiments. It thus seems that item ratings were influenced more by conscious thought than SPL button-press times and changes in pupil size. In all three interactions, individuals with higher political scores (i.e. leaning towards the more progressive side) showed a larger discrepancy between clashing and non-anomalous items (cf. Figs. 2.1d, 2.3c, and 2.4d). As mentioned previously, progressive-leaning individuals rated non-anomalous items better than their more conservative peers, but at the same time rated clashing items *worse*. For socio-cultural clashes, highly progressive individuals rated clashing items about the same as conservative individuals, but they rated non-anomalous items significantly *better* (cf. Fig. 2.4d). These findings are interesting, as intuitively one might expect that progressive raters would generally care less about clashes rooted in established gender stereotypes, and thus rate them better than their conservative peers. Although this finding warrants more research, the findings overall suggest that political views have a significant effect on how listeners evaluate different types of errors and clashes, and that progressive individuals generally have a wider "scale" on which to evaluate utterances. Conservative listeners appear to be more, well, conservative in their judgment, whereas progressive listeners seem to dole out ratings more liberally.

An interaction between **item condition and Extraversion** was found to significantly influence ratings of morpho-syntactic errors and semantic anomalies (Figs. 2.1e and 2.1e). As discussed, for morpho-syntactic errors, ratings were generally lower (across non-anomalous and anomalous items) the more extraverted the listener was, with a steeper slope for anomalous items (cf. Fig. 2.1e). In contrast, for semantic anomalies, highly extraverted listeners rated anomalous items *better* than their less extraverted peers, so that the discrepancy between anomalous and non-anomalous items was *smaller* for highly extraverted listeners (see the size of the "gap" between the slopes in Fig. 2.3b). Note how, as mentioned previously, this means that the slope for semantic anomalies runs *in the opposite direction* compared to that for morpho-syntactic errors (compare the red slopes in Figs. 2.1e and 2.3b). This suggests that the comprehension of morpho-syntactic errors and semantic anomalies is influenced by Extraversion – but in different ways: A highly extraverted person seems to rate morpho-syntactic errors as less acceptable than an introverted person, but at the same time seems to view a *semantic* violation as *less severe* than their introverted counterpart. As discussed previously, this could conceivably stem from extraverted individuals on the one hand being exposed more to atypical stimuli, simply through more social interaction, and extrapolating more from the speaker's native accent on the other (refer back to Section 2.5.2 for details).[3] The conceptual difference between the two – between an error that really can never be considered "correct" when a native speaker produces it, and an anomaly that is odd, but that could, in certain contexts, be acceptable (recall the "the girl comforts the clock" example from Nieuwland and Van Berkum 2006) – appears to be of significance here.

Both Boland and Queen (2016) and this experiment garnered ratings in regards to morpho-syntactic errors – with one big difference: Participants were asked to rate the acceptability of the utterance itself in this ratings experiment, whereas they were asked to rate

[3] Note that, while an assessment of the listener's creativity or imaginative skills was not a part of this research, some prior research has found weak correlations between Extraversion and some (but not all) measures of creativity or imagination (Furnham et al. 2013; Sánchez-Bernardos et al. 2015; Weibel et al. 2018). As such, it is possible that more extraverted individuals engage in more creative, out-of-the-box thinking, rendering semantic anomalies less anomalous.

the suitability as a housemate of the person that made the error in Boland and Queen (2016). Results highlight an interesting difference: While extraverts viewed morpho-syntactic errors as “worse” than introverts did when asked about the utterance itself, they would rate *the person who made the error* as a better housemate than an introvert would. While a detailed exploration of where this disconnect stems from goes beyond the scope of this dissertation, it seems that extraverted individuals, while certainly noticing the error and not approving much of it, are able to set it aside when considering the character of the person who made the error.

An further interesting finding was that the **interaction between Openness and item condition** in the socio-cultural type resembles the interaction with Agreeableness for semantic anomalies much more than it does the interaction with Openness for morpho-syntactic errors - compare Figs. 2.4c and 2.3a on the one hand, and 2.4c and 2.1d on the other. This suggests that Openness has a different effect on the rating of socio-cultural clashes than it has on the rating of morpho-syntactic errors: Whereas for morpho-syntactic errors, more Openness seems to result in a wider “scale” on which to rate items, and hence in a larger difference between correct and erroneous items, more Openness was associated with *higher* ratings for socio-cultural clashes, and thus a smaller difference between clashing and non-clashing items. Interestingly, this finding for socio-cultural clashes meshes with the findings in Boland and Queen (2016), where more open individuals rated the author of an email as a better prospective housemate than their less open counterparts when the email contained typos. As already discussed above, in the context of the interaction between item condition and Extraversion, the difference between the two studies is that this experiment had participants assess the acceptability of the utterance itself, while Boland and Queen (2016) had participants rate the author of the utterance. Much like extraverted individuals in the previous interaction discussed, more open individuals seem to notice morpho-syntactic errors more (potentially due to extrapolating from the speaker’s native accent that morpho-syntactic errors are unlikely to occur), and rate them as not all too acceptable - but they do not seem to let this be reflective of the *character of the person that produced the error*. In that sense, as is also evident from the fact that open individuals rated socio-cultural clashes better than their less open counterparts, more open individuals seem to allow more “leeway” in terms of identity expression, and, very simply, do not judge others as harshly.

It is interesting to note that, while both the ratings of semantic anomalies and socio-cultural clashes were influenced by **speaker gender**, and semantic anomalies by the **listener's gender**, neither of the two gender-related variables had a significant effect on morpho-syntactic errors. This suggests that not all aspects of the listener's and speakers identity influence the processing of an utterance at all times, and in all cases. For example, in this experiment, listener and speaker gender appear to have no influence on the processing of morpho-syntactic errors, arguably the type of clash that draws least on the human experience in the world in its interpretation.

In summary, this chapter showed that the listener's personality, and aspects inferred about the identity of the speaker, influenced the perception of linguistic stimuli even in an off-line ratings experiment. Furthermore, the three different kinds of clashes seem to be influenced by different effects and interactions - beyond a main effect of condition and an interaction between condition and political values, there was very little overlap in extra-linguistic variables across the different models and clash types. This suggests that not all variables affect language comprehension across the board, and that the processing of different stimuli may recruit different comprehension processes and strategies.

We will now move on to the second experiment, which investigated the effect of listener- and speaker-related variables on button-press responses in a self-paced listening study.

Chapter 3

Experiment II: Self-Paced Listening[1]

In this second experimental chapter, we will investigate whether a listener's personality or gender, or the speaker's (inferred) gender, influences the comprehension of anomalous utterances as compared to a non-anomalous baseline in a self-paced listening study. Participants were presented with the same stimuli that were used in Experiment I. Comparing results between these first two studies may give us insight into whether different personality traits affect language processing differently depending on the task at hand, and depending on the type of measurement (acceptability ratings in this chapter, vs. response times in the SPL experiment above).

On-line tasks, such as eye-tracking, self-paced reading (SPR) or listening (SPL), and EEG experiments, provide continuous measurements of the process (Traxler 2014) and can afford great insight into automated language processing. Both self-paced listening and self-paced reading experiments have been found to be well suited to investigate fine-grained comprehension processes (see e.g. De Vincenzi et al. 2003; Roberts 2012; Tokowicz and Warren 2010), with SPR being widely used in e.g. second-language acquisition research (see e.g. Marinis 2003). SPR experiments have been found to be able to detect the exact same effects as an EEG experiment, only a little later (Van Berkum et al. 2005). As the proposed study uses auditory stimuli throughout, this experiment used a timed self-paced listening (rather than reading) paradigm.

[1] Parts of this chapter were presented at the *Alberta Conference on Linguistics* in Calgary, AB, on 29 October 2016; at the *PsychoShorts* conference in Ottawa, ON, on 24/25 February 2017; and at the *CUNY Conference on Human Sentence Processing* in Cambridge, MA, on 30 March - 1 April 2018.

In this paradigm, participants listen to a list of auditory stimuli (for details on the experiment procedure, also see Section 3.3 below). Each stimulus is presented in several chunks, one at a time, rather than as one single unit. Crucially, the audio for the next chunk does not automatically play after the current segment; rather, participants are instructed to press a button as soon as they have made sense of the segment they just heard, which then triggers the next chunk of the stimulus to be played. Response times (RT's), i.e. button-press times, are recorded in milliseconds.

There is comparatively less research using the SPL paradigm as compared to SPR, mainly due to SPL being a younger paradigm (Papadopoulou et al. 2013); however, they both address the same questions and detect similar effects (Marinis 2003; Roberts 2012). As SPL is highly similar to SPR (Papadopoulou et al. 2013, p. 53), but does not require literacy, it is used in research with pre-literate children (Clahsen 2008). Using a self-paced listening paradigm, effects (surfacing as delays in the button-press response) can be detected at different times: Either as immediate effects at the critical word/in the critical region; as spillover effects onto a neighbouring word/region, or at sentence wrap-up (De Vincenzi et al. 2003; Jegerski and VanPatten 2013; Just and Carpenter 1980; Tokowicz and Warren 2010). Prior research reports that morpho-syntactic errors generally result in longer reading times at the critical word, with semantic anomalies being detected later, and also affecting sentence wrap-up negatively (De Vincenzi et al. 2003; Ditman et al. 2007; Just and Carpenter 1980). Response times are expected to be at baseline at sentence wrap-up, or even *faster* than baseline, for morpho-syntactic errors (Ditman et al. 2007; Tokowicz and Warren 2010). Based on existing literature, we thus predicted longer response times for all three types of anomalous statements. For morpho-syntactic errors, the delay was expected to be significant right at the critical segment (see Table 2.1 and Section 3.2 for details on item segmentation), whereas the delay was only expected to surface in the final wrap-up segment for semantic anomalies. No precise predictions could be made regarding socio-cultural clashes; however, a delay was expected somewhere between the critical and wrap-up segments if extra-linguistic information extrapolated from stereotypes is indeed integrated into comprehension rapidly.

Afterwards, the participants' personality and political values were assessed as well, using again the same tests as in the self-paced listening study in Experiment I (refer to Section 2.4 in the previous chapter, and Section 3.4 below). However, there is currently no research regarding how the five personality traits influence SPL response times – recall that, for example in Van den Brink et al. (2010), only an empathy dimension was used, and that Boland and Queen (2016) used an off-line ratings paradigm. Extrapolating from these existing results, we expected to see significant effects of, for example, Openness, Conscientiousness, and Extraversion, even though clear predictions regarding the directionality of the effects could not be made in advance – it is conceivable that high Extraversion, for example, would cause a listener to engage in more anticipation based on stereotypical information, and hence experience more surprisal and longer response times; at the same time, low Extraversion might, due to less exposure to anomalous items, trigger larger processing loads to integrate anomalous information, and hence longer response times.

3.1 Participants

In total, 53 native speakers of English, students recruited from the undergraduate linguistics pool at the University of Alberta, participated in this experiment. Two participants were excluded from analyses as their comprehension question accuracy rate (see Section 3.3 below for details) was well under 80% (72.2% and 69.9%, respectively; $min = 69.9\%, max = 100\%, mean = 96.7\%, median = 96.9\%, SD = 4.5\%$), and their attention to the experiment or the proper execution of the task could hence not be guaranteed. While participants who self-reported a history of psychological or neurological disorders, or hearing loss, were able to participate in the study, their data ($n = 6$ for disorders, $n = 1$ for hearing loss, $n = 1$ for both) was excluded from analyses as conditions like sociopathy, psychopathy, and aphantasia have been found to inhibit empathetic behaviour (Zeman et al. 2015), and hearing loss could prevent proper exposure to the auditory stimuli. The data from 43 native speakers of English (males/females = 21/22; age = 17–25; mean = 19.9 years) was hence used for the analyses below.

3.2 Materials

This experiment re-used the stimuli from Experiment I (see Table 2.1 for the template, and Appendix A for the full list of items; for details, refer back to Section 2.2), both for reasons of comparability, but also with the goal to provide average item ratings that could be included in the statistical models as a numerical predictor. The only difference to Experiment I was that stimuli were not presented as a single auditory unit, but chopped into five segments, or regions, in Praat (Boersma and Weenink 2016). The "critical region," the segment of special interest where all manipulations occurred, contained the verb and the object (refer to Table 2.1); response times should be analyzed for this region rather than the two constituents separately, as morpho-syntactic errors cannot be formed on anything else but the verb, but socio-cultural clashes and semantic anomalies by definition require an object to unfold. Fusing the verb and object into a "critical region" might sacrifice some accuracy, but makes response times comparable across all three conditions. The critical region was separated from the wrap-up region, the second segment of special interest, by a post-critical segment of minimally one syllable (but generally two or three; mean syllable count 2.26) in length (Braze et al. 2002; De Vincenzi et al. 2003; Jegerski and VanPatten 2013) to clearly distinguish immediate effects from wrap-up effects. The critical region was controlled for frequency using the *Corpus of Contemporary American English* (COCA, Davies 2008; see also De Vincenzi et al. 2003; Ni et al. 1998).

All items were rated for acceptability in a separate experiment (refer back to Chapter 2) by a separate set of participants, especially as semantic anomalies and socio-cultural clashes cannot be considered "erroneous" in the same way as morpho-syntactic violations. A pre-ratings experiment is able to catch the inherent gradient perception that comes with anomalies and clashes far better than a binary correct/erroneous distinction would. The average per-item ratings resulting from the separate ratings experiment – instead of a categorical correct/erroneous distinction – were fed into the statistical models as a numerical predictor.

3.3 Procedure

After a short briefing, participants were seated at a desktop computer and asked to wear the headphones provided. They were then presented with one list of items, coded in E-Prime (Psychology Software Tools Inc. 2012), and asked to press the space bar on the keyboard as soon as they had made sense of an auditory segment. The first three items were practice items, and participants were given the chance to ask questions after the practice section. In each trial, the audio of the first segment began playing 100ms after a fixation cross was presented in the centre of the screen. The fixation cross remained on the screen until after the final segment of each item was played, and until the screen showed either "No question. Press SPACE to move on to the next segment," or a simple comprehension question (after approximately 30% of items.)

Comprehension questions were simple, non-anomalous *yes/no* questions in line with well-established world knowledge, such as "Do giraffes have long necks?" after the unrelated filler item "Giraffes always have very long necks," to check for both attention to the experiment, and comprehension of the auditory stimuli that were presented (see e.g. De Vincenzi et al. 2003; Hanulíková et al. 2012).

3.4 Post-Tests

All participants completed the same political questionnaire, Big Five personality assessment, and language background questionnaire as in Experiment I (for details, see Section 2.4 and Appendix B), to maximize comparability.

3.5 Results

All results reported below were, just as in Experiment I, obtained through fitting linear mixed effects models (*LMERs*) using the `lme4` (Bates et al. 2015, version 1.1-21) and `lmerTest` (Kuznetsova et al. 2017, version 3.1-0) packages in `R` (R Core Team 2019, version 3.5.3), with response times (*RT's*; square root transformed) as the dependent variable. For further analysis, reporting, and visualization, the `effects` (Fox and Hong 2009, version 4.1-0), `stargazer` (Hlavac 2018, version 5.2.2), `ggplot2` (Wickham 2016, version 3.1.0), and `MuMIn` (Bartoń 2018, version 1.42.1) packages were used.

Of particular interest are the critical, post-critical, and wrap-up segments, as these are the segments where prior literature has found significant delays in response to the respective errors: Morpho-syntactic errors generally result in longer reading times at the critical word, with semantic/pragmatic errors being detected later, and also affecting sentence wrap-up negatively (De Vincenzi et al. 2003; Ditman et al. 2007; Just and Carpenter 1980).

All LMERs reported below include random intercepts for participant and item. By-item and personality trait random slopes were tested, but resulted in near-singular/overfitted models, so that they could not be used to successfully model the data. Models were fitted and selected using a backwards step-wise elimination procedure, comparing each iteration via ANOVAs and the Akaike Information Criterion (AIC). In all models reported below, the control variables of response time to the previous segment, and the participant's progress in the experiment session ("trial effect") were found to significantly influence response times: A longer response time to the previous segment was correlated with a longer response time to the segment in question, and response times generally became shorter the further a participant had progressed in the experiment. Note that the duration of each auditory segment was tested as a numerical predictor as well, however it was found to be (1) correlated strongly with response time to the previous segment, and (2) less good of a predictor than said response time to the previous segment, so that it was not included in the final model structures.

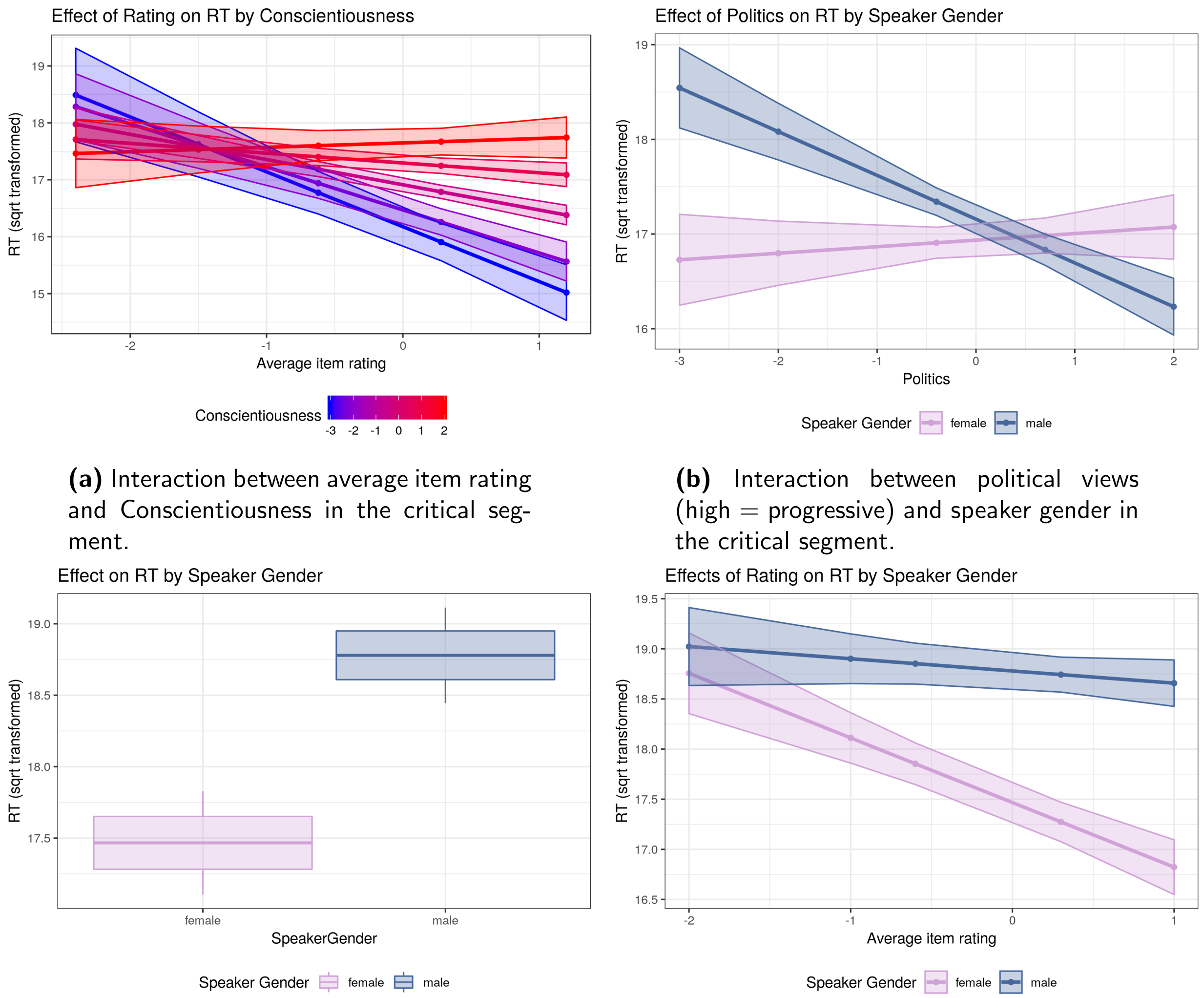
3.5.1 Morpho-Syntactic Errors

For morpho-syntactic errors, recall that a very clear distinction between correct and erroneous items emerged from the ratings experiment, where participants had rated each stimulus on a four-point Likert scale, from zero ("not acceptable") to three ("fully acceptable"). There was barely any overlap in ratings between correct ($median = 2.4; mean = 2.3; SD = 0.4$) and erroneous sentences ($median = 1.1; mean = 1.0; SD = 0.2$; cf. Fig. 2.1a in Chapter 2), so that average item ratings can be read as an accurate reflection of the correctness of an item.

Critical segment: The model with the best fit ($AIC = 11749.51$, marginal $R^2 = 0.10$, conditional $R^2 = 0.30$; for the full model output, see Table 3.1) found a significant **main effect of item rating**, where participants' responses were more delayed the worse an item was rated. This main effect of rating is in line with previous literature, where morpho-syntactic errors have been found to cause a delay in response times right at the critical segment (De Vincenzi et al. 2003; Ditman et al. 2007). Significant interactions were observed between **item rating and Conscientiousness**, where low Conscientiousness scores meant a larger discrepancy in RT's between correct and erroneous items: Responses to correct items were faster, and responses to erroneous items were *delayed* compared to highly conscientious individuals. In contrast, listeners with high Conscientiousness scores showed no effect of rating, i.e. no discrepancy between correct and erroneous items (cf. Fig. 3.1a.) This suggests that highly conscientious individuals generally responded more slowly to correct stimuli than their less conscientious counterparts, and were affected much less, relatively speaking, by errors. This could be explained as highly conscientious individuals wanting to "do it right"

	Effects on RT		
	Critical	Post-Critical	Wrap-Up
(Intercept)	16.936 p = 0.000***	21.326 p = 0.000***	17.467 p = 0.000***
RT to previous segment	1.331 p = 0.000***	.659 p = 0.000***	1.336 p = 0.000***
Progress in experiment	−.513 p = 0.00000***	−.619 p = 0.000***	−1.132 p = 0.000***
Item rating	−.339 p = .002**		−.645 p = .0004***
Speaker gender			1.312 p = 0.00000***
Item rating : Consc.	.208 p = .029*		
Political values : Speaker gender	−.531 p = .005**		
Item rating : Speaker gender			.523 p = .025*

Table 3.1: LMER output for the three segments of interest in the morpho-syntactic error condition. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Note that RT's were square root transformed, and all numerical predictors were scaled and centered.



(a) Interaction between average item rating and Conscientiousness in the critical segment.

(b) Interaction between political views (high = progressive) and speaker gender in the critical segment.

(c) Main effect of speaker gender in the wrap-up segment.

(d) Interaction between average item rating and speaker gender in the wrap-up segment.

Figure 3.1: Visualizations of effects on RT's to morpho-syntactic errors. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Ribbons in line plots indicate the default standard error as calculated via the `effects` package. Top panel: critical segment; bottom panel: wrap-up segment.

- they want to make sure they really have made sense of the segment they heard, thus eradicating any differences between correct and erroneous segments. Furthermore, **speaker gender interacted significantly with political value scores**: For more conservative participants (i.e. those participants with lower scores on the political values scale), RT's were much slower when the speaker was male. For more progressive participants, i.e. those with higher scores on the political values scale, the difference between speaker genders was much less strong, and RT's were slower when the speaker was female (cf. Fig. 3.1b.)

Post-critical segment: No significant main effect or interaction (beyond the two control variables) was found (cf. Table 3.1; $AIC = 11,144.3$, marginal $R^2 = 0.05$, conditional $R^2 = 0.42$.) This is somewhat in line with reports in the literature, which suggest that effects of morpho-syntactic errors are observed quickly, but also die off quickly (De Vincenzi et al. 2003; Ditman et al. 2007); however, as is described below, the main effect of rating returned in the wrap-up segment, so that the effect seemed to not so much "die off" in the post-critical segment, as vanish temporarily.

Wrap-up segment: Prior literature suggests that morphological errors either have no effect on RT's in the wrap-up region, or result in *faster* RT's here (Tokowicz and Warren 2010). This is not what was found in this study; the best model ($AIC = 14,439.2$, marginal $R^2 = 0.10$, conditional $R^2 = 0.35$; for the full model output, see Table 3.1) suggests that there is a **main effect of item rating**, where RT's are slower for items with a low average rating (i.e. erroneous sentences), and faster for items with better ratings, with the effect being *stronger* than in the critical segment. This suggests that the statistical measures used in previous literature, such as ANOVA's, may not have been fine-grained enough to detect the effect of morpho-syntactic errors on sentence wrap-up. The model also found a significant **main effect of speaker gender**, whereby items spoken by a male speaker generally elicited slower RT's compared to those spoken by a female speaker (cf. Fig. 3.1c). An interaction between speaker gender and item rating was observed as well, where again items produced by a female speaker elicited a comparatively stronger reaction compared to a male speaker: For low-rated (i.e. incorrect) items, RT's were rather similar between speaker genders, but RT's then followed a steep slope towards much faster RT's the better the rating for the female speaker, whereas the slope is much shallower for the male speaker (cf. Fig. 3.1d). This effect could very well be a reflection of female speakers being generally more intelligible than male speakers (Bradlow et al. 1996); however, it is also possible that the effect has to do with idiosyncratic speech styles of the two speakers. We will return to this in the discussion below.

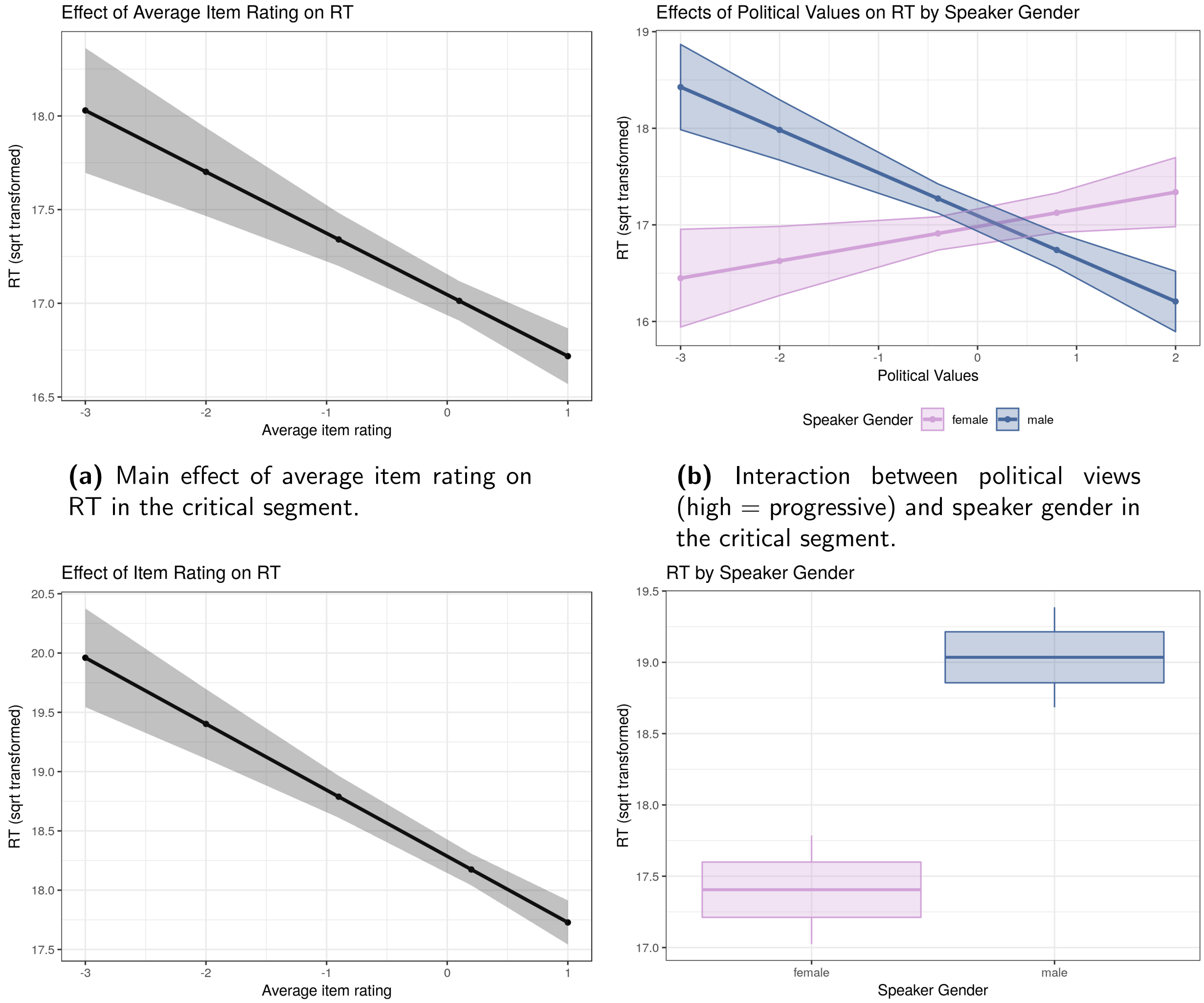
3.5.2 Semantic Anomalies

For the semantic anomaly type, again a clear distinction in average item ratings emerged, as per the ratings experiment. There was a little more overlap in ratings between non-anomalous ($median = 2.4; mean = 2.3; SD = 0.4$) and anomalous sentences ($median = 1.1; mean = 1.3; SD = 0.7$) than for the morpho-syntactic condition (cf. Fig. 2.2a in Chapter 2), but average acceptability ratings still formed two very different distributions, so that they can still be read as clear reflections of item condition, while at the same time reflecting the inherent gradedness in the perception of this kind of anomaly.

		Effects on RT	
	Critical	Post-Critical	Wrap-Up
(Intercept)	16.983 p = 0.000***	21.442 p = 0.000***	17.406 p = 0.000***
RT to previous segment	1.428 p = 0.000***	.675 p = 0.000***	1.395 p = 0.000***
Progress in experiment	−.494 p = 0.00000***	−.665 p = 0.000***	−1.147 p = 0.000***
Item rating	−.328 p = .004**		−.558 p = .00004***
Speaker gender			1.630 p = 0.000***
Political values : Speaker gender	−.622 p = .002**		

Table 3.2: LMER output for the three segments of interest in the semantic anomaly condition. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Note that RT's were square root transformed, and all numerical predictors were scaled and centered.

Critical segment: The best model ($AIC = 10,758.96$, marginal $R^2 = 0.11$, conditional $R^2 = 0.29$; for the full model output, see Table 3.2) found a **main effect of average item rating** (cf. Fig. 3.2a). This is not in accordance with prior literature, which suggests that delays caused by semantic anomalies have a later onset, and only begin surfacing in the post-



(a) Main effect of average item rating on RT in the critical segment.

(b) Interaction between political views (high = progressive) and speaker gender in the critical segment.

(c) Main effect of average item rating on RT in the wrap-up segment.

(d) Main effect of speaker gender in the wrap-up segment.

Figure 3.2: Visualizations of effects on RT's to semantic anomalies. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Ribbons in line plots indicate the default standard error as calculated via the `effects` package. Top panel: critical segment; bottom panel: wrap-up segment.

critical or wrap-up segments (De Vincenzi et al. 2003; Ditman et al. 2007). Furthermore, an interaction was found between an individual's **political values and speaker gender**, with RT's trending in opposite directions for the two speaker genders: Individuals with low political value scores, i.e. leaning to the conservative side, showed faster RT's when the speaker was female, but *slower* RT's when the speaker was male (cf. Fig. 3.2b). This effect is reminiscent of the interaction found in for morpho-syntactic items discussed above and will be discussed below.

Post-critical segment: There was no main effect of average item rating in the best model ($AIC = 10,079.19$, marginal $R^2 = 0.05$, conditional $R^2 = 0.44$; for the full model output, see Table 3.2) - much like for morpho-syntactic errors, the main effect of error condition experienced a "hiatus" in the post-critical segment. This is not in accordance with prior literature, where morpho-syntactic and semantic anomalies exhibited rather distinct delay patterns (De Vincenzi et al. 2003; Ditman et al. 2007). Interestingly, no personality predictors interacted with item rating or speaker gender in this post-critical segment, much like in the morpho-syntactic condition discussed previously.

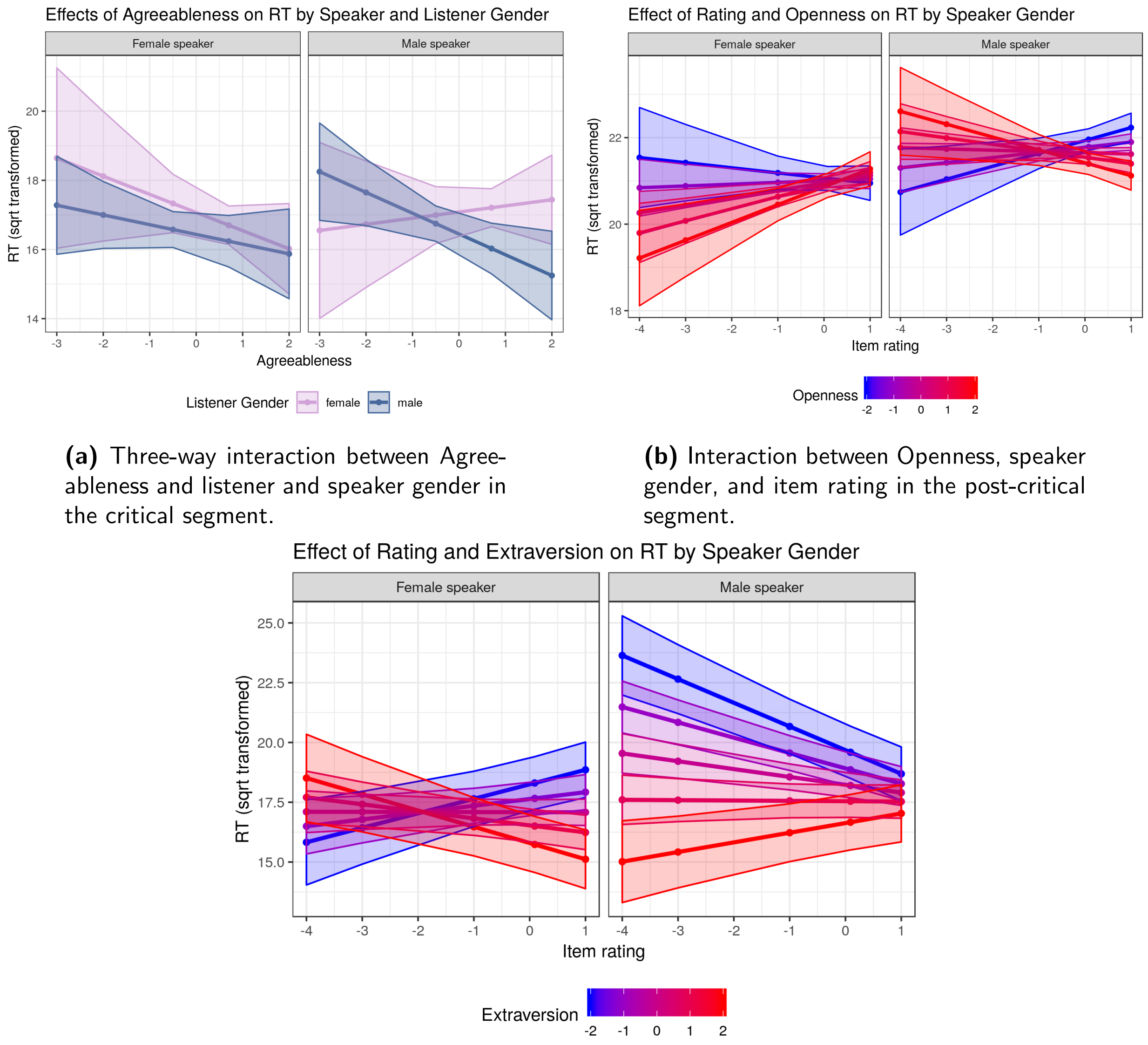
Wrap-up segment: The best model found **main effects of speaker gender and average item rating** ($AIC = 13,197.62$, marginal $R^2 = 0.11$, conditional $R^2 = 0.35$; for the full model output, see Table 3.2), where a higher rating meant faster response times (cf. Fig. 3.2c). This is in line with effects found in previous literature, where semantic effects were predominantly found at sentence wrap-up. Overall, much slower RT's were found in response to stimuli spoken by a male speaker (cf. Fig. 3.2d), which is highly reminiscent of the same main effect in the final segment of the morpho-syntactic error type. No significant interactions, or effects of listener personality, were found in this segment.

3.5.3 Socio-Cultural Clashes

Recall that, in the socio-cultural condition, as expected, the difference in mean item ratings was much less clear-cut than for the previous two item types. Non-anomalous sentences ($median = 2.4; mean = 2.3; SD = 0.4$) were rated only slightly better on average than clashing sentences ($median = 2.2; mean = 2.1; SD = 0.3$; cf. Fig. 2.4a). As predicted, the acceptability of an item was captured in a more fine-grained manner using ratings rather than a simple factorial distinction.

	Effects on RT		
	Critical	Post-Critical	Wrap-Up
(Intercept)	17.069 p = 0.000***	20.965 p = 0.000***	17.156 p = 0.000***
RT to previous segment	1.320 p = 0.000***	.663 p = 0.000***	1.319 p = 0.000***
Progress in experiment	−.516 p = 0.000***	−.531 p = 0.000***	−1.107 p = 0.000***
Speaker gender		.710 p = 0.00000***	1.131 p = 0.000***
Agr. : Speaker : Listener gender	−1.022 p = .019*		
Open. : Rating : Speaker gender		−.282 p = .030*	
Extr. : Rating : Speaker gender			.670 p = .001***

Table 3.3: LMER output for the three segments of interest in the socio-cultural clash condition. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Note that RT's were square root transformed, and all numerical predictors were scaled and centered. Only predictors significant at the .03 level are shown.



(a) Three-way interaction between Agreeableness and listener and speaker gender in the critical segment.

(b) Interaction between Openness, speaker gender, and item rating in the post-critical segment.

(c) Interaction between Extraversion, item rating, and speaker gender in the wrap-up segment..

Figure 3.3: Visualizations of effects on RT's to socio-cultural clashes. Individual difference variables, i.e. political values and personality traits, all pertain to the listener. Ribbons in line plots indicate the default standard error as calculated via the `effects` package.

Critical segment: The best model ($AIC = 14,788.49$, marginal $R^2 = 0.11$, conditional $R^2 = 0.34$; for the full model output, see Table 3.3) found a **significant interaction between speaker and listener gender and Agreeableness**. The presence of a personality predictor, Agreeableness, among the significant predictors in this critical segment suggests that the processing of socio-cultural clashes is influenced by the listener's identity early on. Additionally, no interactions with listener gender were found to be significant in any of the models for morpho-syntactic or semantic anomalies, which suggests that listener gender plays a role only in the processing of stereotype-related clashes. The visualization of the three-way interaction between Agreeableness and listener and speaker gender (Fig. 3.3a) reveals that, when the speaker is female, both male and female listeners respond similarly to stimuli: Less agreeable listeners are slightly delayed compared to their more agreeable peers, with less-agreeable female listeners being comparatively more delayed. However, a very different picture emerges for stimuli spoken by a male speaker: If the listener is female, RT's are now *faster* the less agreeable the listener is. Responses for male listeners are mirrored, with less agreeable male listeners showing the largest delays. Very generally speaking, this suggests that gender plays a role in language comprehension *over and above* personality traits, interacting with the latter. This is an interesting effect, especially following the findings in Gruber and Gaebelein (1979), where male speakers are generally given more attention, and Orlob (2017), who found that men listen more to other men. If RT's are to be taken as a representation of attention, then the same appears to be true in this experiment, at least for men who are less agreeable than average.

While there was no main effect of item rating, the interaction between between speaker and listener gender and Agreeableness suggests that the processing of socio-cultural clashes is affected early on by different facets of the listener's identity.

Post-critical segment: A **main effect of speaker gender** was found in the best model ($AIC = 13,742.19$, marginal $R^2 = 0.06$, conditional $R^2 = 0.45$; for the full model output, see Table 3.3), whereby RT's were slightly slower when the speaker was male. As mentioned previously, this effect could be related to female speakers generally being more intelligible than their male counterparts (Bradlow et al. 1996), so that male speech would require more resources to unpack, reflected in longer reaction times. Alternatively, the delay could result from allocation of attention: Male speakers have generally been found to be more closely attended to than female speakers, even when the message delivered was the same (Gruber and Gaebelein 1979), which could result in longer response times as well.

Interestingly, the socio-cultural clash type was the only clash type in which this main effect of speaker gender surfaced already in the post-critical region, as opposed to in the wrap-up region. This suggests that, in the processing of socio-cultural clashes, which – in this dissertation – hinge crucially on the speaker's gender, information about the speaker may be recruited earlier than for the other two clash types, thus affecting button press times earlier on.

The significant three-way interaction between **item rating, Openness, and speaker gender** suggests that, for either speaker, RT's were rather similar in response to non-anomalous items (cf. Fig. 3.3b), with RT's only slightly longer when the speaker was male. However, the worse the average item rating, the stronger RT's fanned out, depending on the listener's Openness score. Interestingly, opposite ends of the Openness scale resulted in different directions of the effect for male and female speakers: When the speaker was female, low Openness scores were correlated with delays to clashing items; when the speaker was male, it was *high* Openness values that correlated with a delay. This pattern suggests that highly open listeners are thrown off less by errors when the female speaker produced them, and more so when the speaker was male. This is not what was intuitively expected - based on the fact that traditionally, men are given more attention (Gruber and Gaebelein 1979), are the unmarked gender (see e.g. Tannen 1993), and are generally seen as more "competent" (see e.g. Moss-Racusin et al. 2012; Uhlmann and Cohen 2005), the expectation was that it would be *less* open individuals who would rely more on these traditional stereotypes, and hence be thrown off more by errors that were produced by a male speaker. This finding runs counter to the results in Van den Brink et al. (2010), where high empathy was associated with more anticipation and surprisal. However, it should be noted that, even if the sub-trait of "considering opinions other than my own" may be shared, empathy and Openness describe distinct traits, so that results may not be immediately comparable. However, even if the details are unclear at this time, the interactions discussed again lend support to the notion that personality traits and gender identity complement one another, and that a male listener would not necessarily respond to a stimulus in the same manner as a female listener with the same personality traits.

Wrap-up segment: The best model ($AIC = 17857.37$, marginal $R^2 = 0.12$, conditional $R^2 = 0.36$; for the full model output, see Table 3.3) again found no main effect of average item rating, just like in the first two segments. The absence of such an effect could be due to either the coarseness of the SPL paradigm as compared to e.g. pupillometry, such as will be reported below; or to the much less clear-cut distinction between clashing and non-anomalous items, as compared to the more traditional error types. Just like in the post-critical segment, there was a **main effect of speaker gender** (irrespective of clash condition), whereby items spoken by the male speaker generally elicited longer RT's.

A significant interaction between **item rating, Extraversion, and speaker gender** was observed (cf. Fig. 3.3c): When the speaker was male, RT's showed much greater variance than when the speaker was female, especially for low-rated items (i.e. those containing a socio-cultural clash). A highly introverted listener, when presented with a clash spoken by a male speaker, would show the largest differences in RT's between a clashing and a non-clashing item. When the speaker was female, however, a very different picture emerged, one where RT's were *faster* in response to non-anomalous items for highly extraverted listeners. For introverted listeners, this relationship was mirrored: RT's were faster in response to clashing items, and slower in response to correct items. It has to be noted that this interaction effect is much smaller and less clear when the speaker is female. This suggests that the (perceived) gender of the speaker modulates the processing of socio-cultural clashes. In accordance with research already mentioned previously, suggesting that male speakers are generally given more attention than female speakers (Gruber and Gaebelein 1979; Orlob 2017), the longest RT's were observed for socio-cultural clashes when the speaker was male. This suggests that it is more "jarring" for the listener when men produce a strange (as related to traditional stereotypes) statement, as compared to when women produce a statement of this nature. While an analysis of further interactions, such as between the gender of the speaker and the listener, modulated by personality traits, goes beyond the scope of this research, it could be an interesting avenue for future research. Even though there was no significant main effect of rating (i.e. clash condition), the significant interaction reported above suggests that the sentence wrap-up processing of socio-cultural clashes is affected by extra-linguistic variables.

3.6 Discussion

Summing up the findings for **morpho-syntactic errors**, a main effect of average item rating (i.e. error condition) was found in the critical and the wrap-up segments. The main effect of rating in the wrap-up segment was found to be *stronger* than that in the critical segment, even while the random effects of item and participant were accounted for. This is not entirely in line with what previous literature suggests - namely, that delays caused by morpho-syntactic errors become apparent right at the critical segment, and then die off quickly, or even "flip" in polarity in the wrap-up segment (De Vincenzi et al. 2003; Ditman et al. 2007). However, the method of analysis used in this current SPL experiment, linear mixed effects modelling, differs from the ANOVA's used in the literature in a number of crucial ways. Unlike ANOVA's, linear mixed effects modelling does not lose information in averaging over time ranges, and it can consider the influence of random effects and additional predictor variables parsimoniously. This suggests that ANOVA's may not be fine-grained enough to capture the effects at play at sentence wrap-up, and that linear mixed effects modelling, through inclusion of random factors and other predictor variables, may be able to capture the underlying relationship more accurately. Results further suggest that a listener's personality, error condition, and the speaker's identity influence responses to spoken language in the critical and wrap-up segments, but not in the post-critical segment. It is possible that effects are masked in this middle segment, as the participant's response time to the previous segment was included in the statistical models, and as sentence wrap-up processing hasn't yet begun.

A main effect of average item rating was found in the critical and wrap-up segments for **semantic anomalies**. This is not in line with existing SPL literature, where semantic anomalies were found to *not* cause delays early on, but only during sentence wrap-up and potentially in the post-critical segment (De Vincenzi et al. 2003; Ditman et al. 2007; Tokowicz and Warren 2010). It is interesting to note that the same behaviour was observed for morpho-syntactic errors in this experiment, whereas prior literature found these two error types eliciting slightly different effects. This could potentially be due to the different methods of statistical analysis, as discussed above.

The only individual differences that had an effect on RT's to semantic anomalies were the listener's political values and the gender of the speaker. Interestingly, political value scores were found to be a significant predictor in only the critical segment for both the morpho-syntactic and semantic anomalies, whereas speaker gender was found to be significant in both the critical (in an interaction) and the wrap-up segments (as a main effect and/or interaction). The main effect of the political values score subsides after the critical segment in both cases, whereas the speaker gender effect resurfaces on the wrap-up segment.

Recapping the findings for **socio-cultural clashes**, no main effect of average item rating was found in any of the three segments. This may sound a little surprising at first, as prior ERP literature has found inference-based clashes to affect language comprehension (Van Berkum et al. 2009, 2008); however, with this study using LMER modelling, the significant influence of rating on RT's was captured in significant interactions with several variables pertaining to the listener's and speaker's identity, such as speaker gender, Openness, and Extraversion. This suggests that the processing of socio-cultural clashes is influenced by item rating on the one hand, and speaker- and listener-related variables on the other.

3.6.1 Personality & Gender Effects

With regards to listener-related variables, it was specifically their Conscientiousness, Openness, Extraversion, Agreeableness, and political value scores that contributed to significant interactions. While these are not variables that are generally investigated in linguistic ERP research, Openness and Conscientiousness are among the most salient predictors in Boland and Queen (2016)'s analysis of "typos" and "grammos" (refer back to Section 1.1.2.3).

Comparing the time-course of significant personality effects across the three error types, a few systematic patterns emerge: In the morpho-syntactic type, significant personality interactions were only found in the critical segment, suggesting that effects of personality influence the processing of this type of error early, and then taper off, "making room" for effects of gender. In terms of effects of personality, RT's to the semantic type were only influenced by an interaction of speaker gender with political values, which tapered off before sentence wrap-up. In contrast, interactions with personality traits were found across *all* segments for the socio-cultural clash type; recall that neither of the models for morpho-syntactic and semantic anomalies found any personality effects or interactions in the post-critical and wrap-up segments. Additionally, listener gender only ever affected

RT's in response to socio-cultural clashes. These findings suggest that the comprehension of socio-cultural clashes differs from the processing of traditional errors, taking into account different listener-related variables at different stages of the process. Main effects of and interactions with speaker gender were observed across all three clash types, where items produced by a male speaker predominantly elicited slower RT's compared to those produced by a female speaker (cf. Figs. 3.1c and 3.2d for two examples). As discussed previously, this effect surfaced in the wrap-up segment across all three clash types, and additionally in the post-critical region for socio-cultural clashes, suggesting that features extrapolated from the speaker's voice may be recruited earlier when they are crucially important for the processing of the clash in question. As for the origin of this effect, listeners heard both speakers produce the same ratio of correct and clashing sentences, in all three clash types, during the experiment – it is hence unlikely that listeners would have formed an expectation to hear the male speaker produce only correct utterances and subsequently experienced surprisal, surfacing as a button-press delay, upon encountering an anomalous sentence. Further, the difference in RT's between speaker genders cannot be due to differences in speech rate, which was controlled for via a per-item measurement in a separate model variable. As noted previously, the speaker gender effect could again be a reflection of female speakers generally being more intelligible than males (Bradlow et al. 1996), resulting in longer reaction times. As another alternative, the delay could result from allocation of attention (Gruber and Gaebelein 1979). Specifically regarding the interaction effects with speaker gender, males are considered the unmarked gender (see e.g. Tannen 1993), and are generally seen as more "competent" (see e.g. Moss-Racusin et al. 2012; Uhlmann and Cohen 2005), so that any deviation from the "competent maleness" would be more surprising than a female speaker producing an error. Lastly, the possibility for the effect to stem from idiosyncratic differences between the two speakers cannot be excluded; we will return to this in the General Discussion.

3.6.2 Effect Patterns

Comparing the overall effect patterns between the three clash types and three segments analyzed, morpho-syntactic and semantic anomalies show largely the same effect pattern, except that the processing of morpho-syntactic errors was affected by two additional interactions (rating by Conscientiousness in the critical segment, and rating by speaker gender in the wrap-up segment). There is much less overlap in the effect patterns between socio-cultural clashes and either of the two more traditional error types; it is only a main effect of speaker gender in the wrap-up segment that all three have in common (and that the model for socio-cultural clashes shares with the model for either of the two traditional errors.) Further, the only clash type that was influenced by any extra-linguistic variable or interaction in the post-critical segment was the socio-cultural type; no effects were found there for morpho-syntactic and semantic anomalies. Finally, the processing of the two traditional anomaly types overall seems to be influenced by very different kinds of listener-related variables as compared to socio-cultural clashes: Whereas it was Conscientiousness and political values that affected RT's for those errors - variables that are comparatively more task-related and slightly more abstract than personality traits - socio-cultural clashes were affected by interactions with Openness, Extraversion, and Agreeableness, i.e. variables that relate much more strongly to human interaction, and to engaging with and approaching other humans (cf. Table 2.2).

Summing up, the results obtained in this SPL experiment suggest that the listener's and speaker's identity, including aspects such as personality and gender, indeed modulate RT's to errors and clashes, and that there is no one effect that modulates RT's across the board. Specifically, morpho-syntactic and semantic anomalies seem to be influenced by different variables, and at different stages, than socio-cultural clashes. It is encouraging to note that these different effects could be shown even using a rather coarse paradigm such as self-paced listening. Results will be tied in and discussed with findings from the off-line ratings study and the two pupillometry studies in the General Discussion (Chapter 6).

We will now turn to the two pupillometry experiments, which provide insight into language comprehension that is not mediated by a task, or by conscious actions. Comparing results between these two very different paradigms can be expected to inform which variable is recruited only during conscious processing, and which come into play in subconscious comprehension.

Chapter 4

Experiment III: Pupillometry[1]

As we have seen in the previous two chapters, both the off-line ratings and timed self-paced listening paradigms were able to identify effects of extra-linguistic variables on language comprehension. However, both of these behavioural paradigms involve measures that are under conscious control of the participant – item ratings on the one hand, and button-press responses on the other – and thus may be affected by overt decision-making. For this reason, this chapter, and the one following it, make use of the pupillometry paradigm. In this paradigm, participants' pupil sizes are monitored continuously throughout each trial, in an entirely non-invasive fashion. In contrast to the previous two chapters, where language comprehension was measured either at the end (as in Chapter 2), or at five defined points (as in Chapter 3), the two pupillometry chapters specifically investigate the influence of personality traits and political values on language comprehension *as it happens.*

Beyond responding to ambient light levels, pupil size is considered an indicator of autonomic nervous system activity (Gingras et al. 2015; Partala and Surakka 2003) that is especially responsive to cognitive effort, mental workload, attention, arousal, and affective processing (Beatty 1982; da Silva-Castanheira et al. 2019; Gingras et al. 2015; Goldinger and Papesh 2012; Just and Carpenter 1993; Kahneman and Beatty 1966; Partala and Surakka 2003; Piquado et al. 2010).[2] Changes in pupil size are not thought to indicate one single process or state, but rather a combination of several. While for example affect and mental

[1] Parts of this chapter were presented at the *PDFA/GSA Research Day* in Edmonton, AB, on 24 October 2018, and at the *Linguistics Department's 50th Anniversary Conference* on 13 April 2019.

[2] Note that this means that we cannot distinguish between a significantly larger pupil dilation that is due to enhanced cognitive load, and one that is due to a difference in affective processing, for example; in the remaining parts of this dissertation, "cognitive load" or "cognitive effort" will generally be used to refer to this group of effects.

workload may seem unrelated, the underlying reason for correlated pupillary differences may be resource allocation (Rondeel et al. 2015), or the "intentional attentional engagement" between the individual and the stimulus (Winn et al. 2018). Pupillometry has been found to be an effective measure for linguistic processing, where it is thought to function as an indicator of the intelligibility, complexity, and/or ambiguity of an utterance (Rij et al. 2019; Vogelzang et al. 2016; Winn et al. 2018). Crucially, language comprehension processes can be analyzed in the absence of a task that might directly draw attention to the phenomena under investigation, unmediated by conscious decision-making. Pupillometry has also been found to reliably identify effects of individual differences (Lõo et al. 2016), and to be especially revealing in tasks with low cognitive load (Gingras et al. 2015). Measuring pupil size during a reading task, Just and Carpenter (1993) found that the reading of more complex sentence types was correlated with larger pupil dilations, and concluded that pupil size is an indicator of the "intensity of thought," i.e. the demand placed on cognitive processing by a linguistic stimulus, supporting Kahneman and Beatty (1966)'s early pupillometric research. In an auditory experiment, Vogelzang et al. (2016) found that pupil sizes were larger when listeners came across an ambiguous pronoun, confirming that pupillometry can be an effective tool in measuring linguistic complexity and the processing load it demands from listeners. A detailed overview of further auditory pupillometry research, investigating e.g. the effects of memory load, linguistic complexity, and more, can be found in Zekveld et al. (2018, specifically Table 1). Partala and Surakka (2003) found that pupil size increased with the emotionality of a stimulus, whether loaded positively or negatively, and concluded that the autonomic nervous system responds differently to emotional stimuli than it does to neutral ones. Pupil size also seems to be able to measure responses to non-linguistic stimuli: Gingras et al. (2015) recorded pupil dilations in response to musical excerpts and found that pupil size was correlated with the excerpt's arousal and tension ratings. They also found gender differences, and differences based on how big of a role music played in the listener's life, suggesting that both the qualities of the musical excerpt, and the attitudes of the listener, affected pupil dilation. Attempting to derive a broader conclusion from existing pupillometric research, Rondeel et al. (2015) devised a series of tasks assessing different components of cognitive control. They concluded that, while pupil size has been found in prior research to be an indicator of such varied facets as cognitive load, affect, and reward, the underlying dimension may be general resource allocation. Tying in with research from the bio-medical field, pupils appear to constrict with parasympathetic activity, and dilate with sympathetic activity (Bradley et al. 2008; Rondeel et al. 2015; Steinhauer et al. 2004; Winn et al. 2018). While the detailed physiological mechanisms behind pupil constriction and dilation are beyond the scope of this thesis, these findings lend further support to pupillometry being an adequate and efficient tool to assess online cognitive processing.

As per the literature discussed above, in which anomalous or unusual stimuli were reflected in increased pupil size, we used pupillometry to assess the cognitive load, or effort, associated with a stimulus. We expected a significant increase in relative pupil size for morpho-syntactic errors, semantic anomalies, and socio-cultural clashes, as compared to a non-anomalous baseline. Crucially, we also expected those changes in relative pupil size to be modulated by an individual's personality or political views, especially in the case of socio-cultural clashes: We expected a more conservative outlook to be correlated with a larger spike in pupil sizes after a socio-cultural clash (recall that those relied on inferences made based on traditional gender stereotypes), for example. Conversely, we expected morpho-syntactic errors to be influenced much less by variables pertaining to the internal state of the listener, as those by definition do not draw on clashes with the real world.

4.1 Participants

57 students, recruited from the pool of students enrolled in introductory undergraduate linguistics courses at the University of Alberta, participated in this experiment. Unfortunately a significant amount of data loss occurred, which could be traced back to an equipment issue (that has since been fixed) during data collection. For this reason, trials from several participants had to be removed, so that the data from 33 participants (males/females = 11/22; native/non-native speakers of English = 28/5; age = 18–23; mean = 19.3 years) was used in the analyses below.

4.2 Materials

For consistency, the same materials from Experiments I and II were re-used, in the same format as in the ratings study (see Table 2.1 for the template, and Appendix A for the full list of items; for details, refer back to Sections 2.2 and 3.2).

4.3 Procedure

After introducing participants to the experimental setup, they were seated in an adjustable chair in a dimly lit experiment booth at the *Centre for Comparative Psycholinguistics* at the University of Alberta. Lighting levels were kept constant throughout the experiment, and for all subjects. While the participants' movements were not restricted, they were asked to place their head on a chinrest to provided additional stability and a constant screen-to-eye distance. Participants were then instructed to follow the instructions on the screen to calibrate the eye-tracker, and to complete the experiment. During the experiment, the pupil size of the participant's right eye (cf. Kahneman and Beatty 1966; Porretta and Tucker 2019) was recorded at 250Hz using an *EyeLink 1000* system on a desktop PC.

Each trial began with a one-point drift correct, and, immediately after, the display of a fixation cross at the centre of the screen. Pupil size was recorded from the start of the fixation cross onwards. 2000ms later, the audio stimulus began to play, and pupil size was recorded until 500ms after audio offset. After approximately 30% of trials, participants were presented with a simple comprehension question (the same questions as for the self-paced listening study; for details, refer back to Section 3.3). After an inter-stimulus interval of 3,000ms, to allow pupil dilation to return to baseline,[3] the next trial began. Participants were given longer breaks approximately every thirty-five trials; the length of these longer breaks was entirely up to the participant. The main experiment thus took between 20 and 30 minutes (40, in rare cases) to complete, depending on how long participants chose their breaks to be (and how often the eye-tracker had to be recalibrated, for example due to movement between trials). Participants then moved on to the posts-tests described below.

4.4 Post-Tests

Participants completed the same post-tests as in Experiments I and II: The political questionnaire, the Big Five personality assessment, and the language background questionnaire. For details, refer back to Section 2.4 and Appendix B.

[3] This value was chosen to be longer than 2,000ms to avoid spillover effects into the next trial (Schmidtke 2018) in line with prior research (Bergamin et al. 1998; Crippa et al. 2018), and to not exceed 3,000ms so as not to slow down the pace of the experiment too much, as this may affect results in unintended ways (cf. Papesh and Goldinger 2012; Schmidtke 2018).

4.5 Results

The accuracy rates of answers to comprehension questions were checked for all participants to test for attention, and for whether the participant can be assumed to have understood the stimulus. Accuracy rates were higher than 80% for all participants ($min = 81.5\%, max = 100\%, mean = 93.4\%, median = 92.5\%, SD = 5\%$), so that no data was removed based on a lack of attention or comprehension of the stimuli. The raw pupillometry data was then downsampled to 50 Hz and preprocessed in `R` (R Core Team 2019, version 3.5.3). Blinks were removed semi-automatically, using Jacolien van Rij's `removeBlinks()` function and visual inspection. All timestamps were time-locked to the onset of the respective clash or anomaly. Baseline pupil sizes were calculated per participant per trial, and outliers further than 2.5 SD's from the respective baseline (around 3% of data points) were removed.

All results reported below were obtained through generalized additive mixed effects modelling (*GAM modelling*, or *GAMM*) using the `mgcv` (Wood 2011, version 1.8-28) and `itsadug` (van Rij et al. 2016, version 2.3) packages in `R`, with relative pupil size being the dependent variable. All models included a random smooth for participant by time, and a random intercept by item to account for individual differences within the stimuli, and for random variance between participants beyond the factors of interest. This makes the analyses markedly different from e.g. the ANOVA's in Grey and Van Hell (2017), Hanulíková et al. (2012), Van Berkum et al. (2008), and Van den Brink et al. (2010). GAM modelling is extremely well suited to time-series research, such as pupillometry, as it is able to capture non-linear interactions between continuous predictors, and as it allows to control for random participant and item effects. Of special importance for the analysis of pupillometry studies is that GAMMs can comfortably model time-series data without losing information in time-binning or averaging, and that it does not assume linear relationships, an assumption that is often unwarranted (Rij et al. 2019; Tremblay and Newman 2015). GAM modelling has been used successfully at our lab, the *Centre for Comparative Psycholinguistics*, in the recent past: Non-linear effects were found for listener experience and the perception of foreign accents (Porretta et al. 2017, 2016), and for pupil size in a naming task (Lõo et al. 2016).

All models were fitted using a forwards step-wise selection procedure. The inclusion of variables was evaluated using a combination of a χ^2 test of REML scores via the `compareML()` function, visual inspection, and the estimated p-value of the smooth parameter via the `report_stats()` function (see e.g. Mukai et al. 2018; Porretta and Tucker 2019; Rij et al. 2019). All numerical predictors were scaled and centered to avoid unintended effects of different orders of magnitude between predictors. Of special interest for this dissertation are three-way interactions between an extra-linguistic variable (such as a personality trait or an individual's political leaning), time since clash onset, and average item rating, which will be visualized using three-dimensional surface plots (explained in detail in Section 4.5.1 and Fig. 4.1c below). Note that separate models were fitted for each individual difference variable, so as to not over-complicate each GAMM; however, each individual predictor that was found to be significant was then fed into a GAMM together with each of the other significant predictors, to test if the effects remained. So, for example, if Openness and political values surfaced as significant predictors in separate models, an additional GAMM was fitted with *both* Openness and political values as predictors, to check that the effects did not cancel each other out. All effects reported below remained in tests of this kind. None of the models reported below found any significant main effects of speaker gender, listener gender, or native speaker status on pupil size.

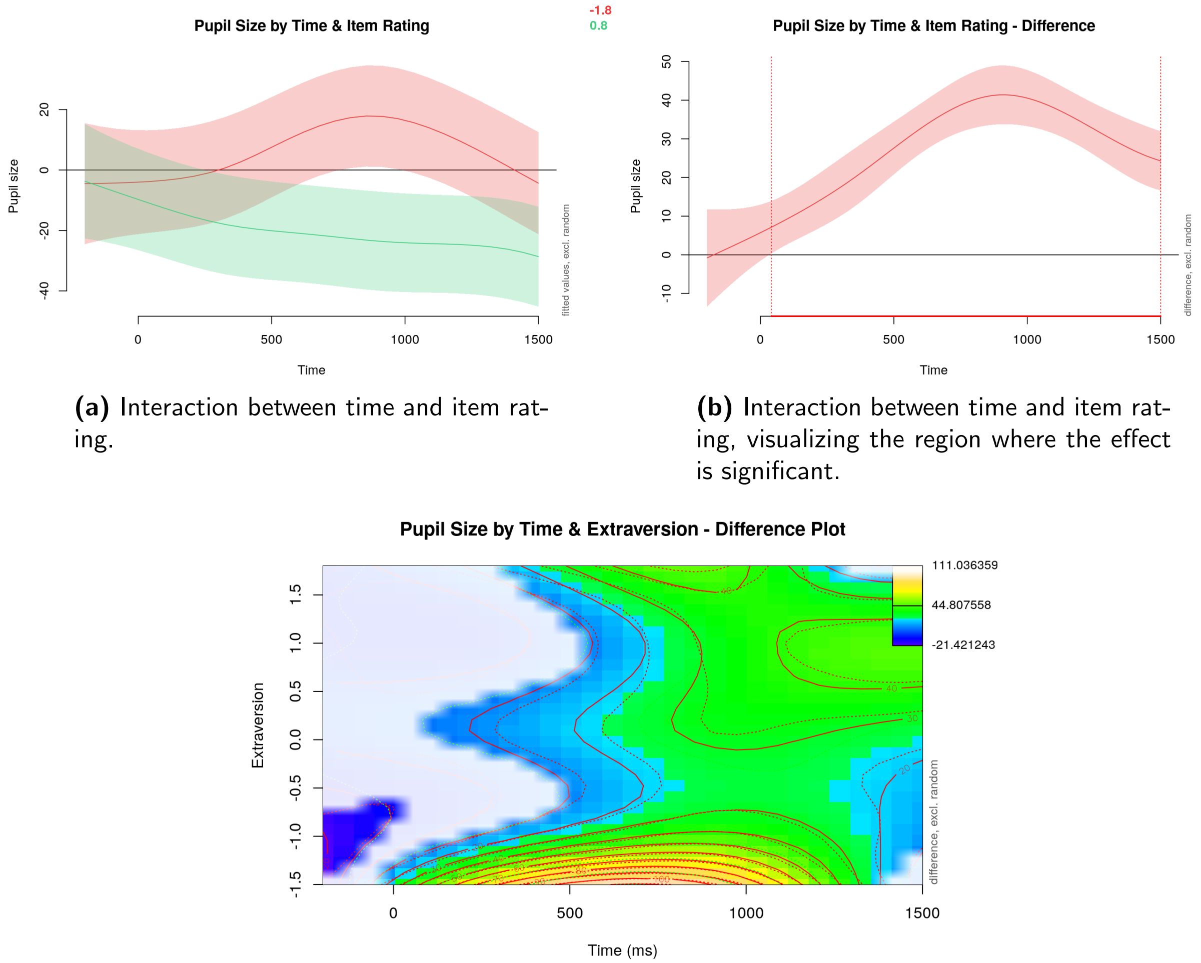
4.5.1 Morpho-Syntactic Errors

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	9.3027	5.5454	1.6776	0.0934
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	4.7213	5.3223	10.5232	< 0.0001
Item rating	8.2791	8.8429	53.9727	< 0.0001
Extraversion	2.3694	2.3857	2.1511	0.0827
Time : rating	14.8230	15.7748	32.1531	< 0.0001
Extraversion : time	5.7483	6.2207	1.1036	0.3765
Extraversion : rating	15.7902	15.9923	48.4939	< 0.0001
Extr. : time : rating	54.7942	60.9302	12.2323	< 0.0001
Random structure				
Participant : time	202.9634	295.0000	40.2479	< 0.0001
Item	99.6157	101.0000	66.2952	< 0.0001

Table 4.1: Output of the best GAM model for morpho-syntactic errors, with the listener's Extraversion as the extra-linguistic predictor.

In the best model for morpho-syntactic errors (pupil size samples $n = 202,436$; see also Table 4.1), a significant interaction between item condition (i.e. average item rating) and time was found: As expected, pupil sizes increased significantly shortly after listeners encountered a morpho-syntactic error (cf. Figs. 4.1a and 4.1b). Of special interest for the research questions in this dissertation, a significant three-way interaction was found between the listener's **Extraversion scores, average item rating, and time.** This relationship is visualized in Fig. 4.1c, where the x-axis shows the time since clash onset (in ms), the y-axis shows the listener's Extraversion score (normalized and centered), and colours denote the relative change in pupil size between a stimulus containing an error, and one that does not. This plot, like all surface plots in the two pupillometry chapters, can be interpreted like a beach landscape: A blue colour represents a smaller change in pupil size, whereas a yellow or orange colour denotes a larger change in relative pupil size (when participants would experience a much larger pupil size when encountering an error); also note the gradient scale in the top-right corner of these surface plots. Areas not significantly different from zero are shaded white. Fig. 4.1c thus shows that less extraverted listeners (i.e. introverted individuals, visualized as smaller values on the y-axis) experienced a much larger change in pupil size over time than did more extraverted listeners – compare the yellow/orange colours near the bottom of the image with the blue and green colours near the top. It thus appears that morpho-syntactic errors, which openly violate established grammatical "conventions," are associated with a larger processing load and mental effort for less extraverted individuals. Interestingly, while a similar effect was found in the ratings study in response to semantic anomalies, where introverted individuals rated anomalous sentences worse than their extraverted counterparts, the opposite was found for morpho-syntactic errors. So, whereas introverted listeners seem to experience a larger cognitive load, as per increased pupil size,[4] when encountering a morpho-syntactic error (cf. Fig. 4.1c), they rated the utterance containing the error as more acceptable than did their extraverted counterparts in an off-line behavioural ratings study (cf. Fig. 2.1e). We will return to a broader discussion of this – seemingly inconsistent – effect of the listener's Extraversion below.

[4] Recall our earlier discussion, at the beginning of this chapter, regarding terminology; "cognitive load," "cognitive effort," and related phrases are used as a stand-in for the group of effects that can be detected using the pupillometry paradigm.



(a) Interaction between time and item rating.

(b) Interaction between time and item rating, visualizing the region where the effect is significant.

(c) Visualization of the three-way interaction between time, item rating, and the listener's Extraversion, with pupil sizes on the z-axis (colour scale).

Figure 4.1: Visualizations of the GAM model for morpho-syntactic errors ($AIC = 2,626,638$) that uses the listener's Extraversion as the extra-linguistic predictor. For details on the surface plot (bottom), and all others of its kind in this dissertation, please refer to the body text in Section 4.5.1.

4.5.2 Semantic Anomalies

In the modelling of semantic anomalies, a significant increase in pupil size was found for worse-rated items (i.e. items containing a semantic anomaly) as compared to non-anomalous items, starting around 400ms after clash onset (cf. Figs. 4.2a and 4.2b). Furthermore, **four extra-linguistic predictors were found to interact significantly with item rating and time** ($n = 182,342$ for all models; see Tables 4.2, 4.3, 4.4, and 4.5 for model summaries); we will now discuss these four three-way interaction effects in turn.

In the interaction with the listener's Openness score, more open individuals were found to experience a larger increase in pupil size when encountering a semantic anomaly. This result is not intuitively accessible, with higher Openness scores generally being associated with greater curiosity, creativity, and unconventionality (cf. Table 2.2). The result could suggest that, akin to a good mood or higher empathy (Havas et al. 2007; Van Berkum et al. 2013; Van den Brink et al. 2010; Zadra and Clore 2011; also refer back to Sections 1.1.2.2 and 1.1.2.3), higher Openness may result in more anticipation as to how the utterance might continue, and thus resulting in more surprisal when a semantic anomaly is encountered. However, the listener's Openness score was not found to be a significant predictor in the modelling of responses to semantic anomalies in any of the other three experiments, so that such a generalization cannot be made with absolute certainty. Interestingly though, the listener's Openness was found to be a significant predictor in the modelling of item ratings in response to morpho-syntactic errors and socio-cultural clashes (refer back to Section 2.5): With more open individuals experiencing a larger change in pupil size in response to semantic anomalies, the effect found in this pupillometry study is similar to that observed for item ratings of *morpho-syntactic errors* (refer back to Fig. 2.1c), but *opposite* to the effect found for socio-cultural clashes (refer back to Fig. 2.4c). This suggests that the overt, conscious rating of items is influenced differently by aspects of the listener's personality than is the immediate, subconscious comprehension of an utterance.

In the interaction with the listener's Extraversion score, introverted listeners showed the largest increase in pupil size in response to semantic anomalies, whereas no significant difference was found in pupil size between correct and clashing items for more extraverted listeners (cf. Fig. 4.2d). This effect is highly similar to that discussed just above, in response to morpho-syntactic errors (compare Fig. 4.2d to Fig. 4.1c in the previous section), and

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	5.7743	5.0764	1.1375	0.2553
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	4.2703	4.8604	7.1400	< 0.0001
Item rating	8.8079	8.9799	55.9606	< 0.0001
Openness	1.0024	1.0025	15.1125	0.0001
Time : rating	14.1234	15.4712	9.2709	< 0.0001
Openness : time	5.0515	5.5619	4.6529	0.0002
Openness : rating	15.9094	15.9982	116.4555	< 0.0001
Openn. : time : rating	57.7458	62.2800	13.8088	< 0.0001
Random structure				
Participant : time	193.6311	295.0000	51.6121	< 0.0001
Item	99.3504	101.0000	67.2006	< 0.0001

Table 4.2: Output of the GAMM for semantic anomalies that uses the listener's Openness as the extra-linguistic predictor ($AIC = 2,363,676$).

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	10.6892	4.9740	2.1490	0.0316
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	4.4838	5.0665	8.4370	< 0.0001
Item rating	8.8316	8.9832	69.2381	< 0.0001
Extraversion	2.2579	2.2859	1.8116	0.1189
Time : rating	13.1506	14.9682	7.1258	< 0.0001
Extraversion : time	5.1978	5.6982	0.6662	0.6710
Extraversion : rating	15.8437	15.9955	93.0960	< 0.0001
Extr. : time : rating	58.0190	62.6032	13.9491	< 0.0001
Random structure				
Participant : time	200.1982	295.0000	26.1647	< 0.0001
Item	99.3285	101.0000	66.0347	< 0.0001

Table 4.3: Output of the GAMM for semantic anomalies that uses the listener's Extraversion as the extra-linguistic predictor ($AIC = 2,364,088$).

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	7.9429	5.0269	1.5801	0.1141
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	4.3695	4.9411	7.8838	< 0.0001
Item rating	8.8178	8.9810	50.4633	< 0.0001
Neuroticism	1.0031	1.0032	1.2973	0.2548
Time : rating	12.3313	14.4312	4.8942	< 0.0001
Neuroticism : time	5.6721	6.2477	2.0001	0.0619
Neuroticism : rating	15.2422	15.9050	29.7843	< 0.0001
Neur. : time : rating	56.5601	61.9409	10.9368	< 0.0001
Random structure				
Participant : time	199.3599	295.0000	36.0827	< 0.0001
Item	99.3231	101.0000	65.8546	< 0.0001

Table 4.4: Output of the GAMM for semantic anomalies that uses the listener's Neuroticism as the extra-linguistic predictor ($AIC = 2,365,240$).

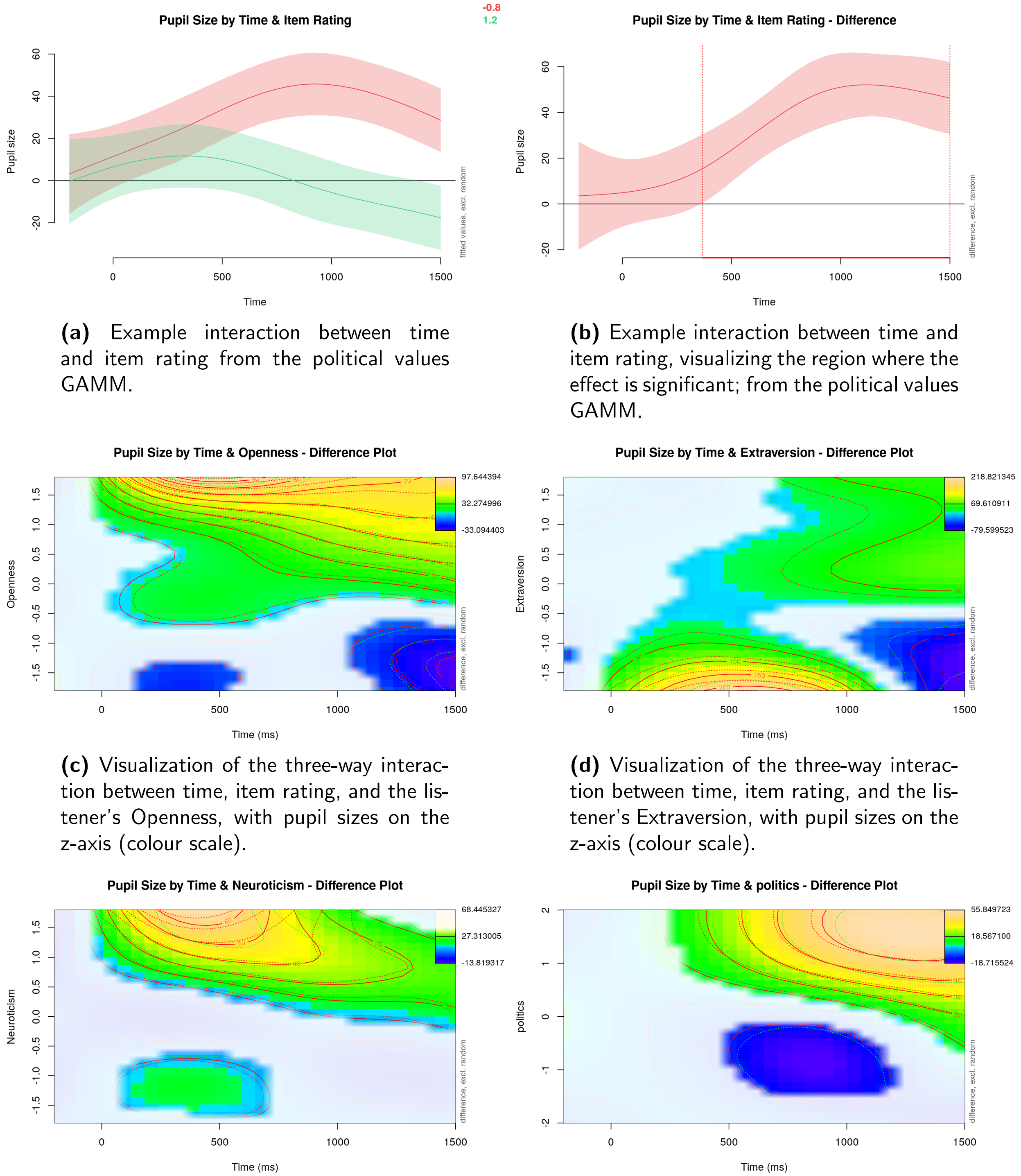
Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	9.0040	5.0395	1.7867	0.0740
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	4.3830	4.9497	7.8465	< 0.0001
Item rating	8.8241	8.9821	51.0488	< 0.0001
Political values	1.0016	1.0018	0.0034	0.9534
Time : rating	12.6821	14.6541	5.9445	< 0.0001
Political values : time	1.6917	1.7873	0.6310	0.5582
Political values : rating	15.4876	15.9534	71.7417	< 0.0001
Pol. : time : rating	55.0848	60.9047	14.0176	< 0.0001
Random structure				
Participant : time	203.9973	295.0000	37.9288	< 0.0001
Item	99.3472	101.0000	66.4373	< 0.0001

Table 4.5: Output of the GAMM for semantic anomalies that uses the listener's political values as the extra-linguistic predictor ($AIC = 2,364,422$).

aligns with the effect of Extraversion on item ratings for semantic anomalies (cf. Section 2.5.2), where introverted listeners rated semantic anomalies worse than their extraverted peers. As already discussed in Section 2.5.2 of the ratings study, a possible explanation for this effect could be that introverted listeners may have had less exposure to this kind of clash, or to unusual statements in general, simply by either socializing less than their extraverted peers – or by attending less diverse social events.

In the interaction with Neuroticism, as is intuitively expected, more neurotic listeners, i.e. those that are characterized as more easily upset and less calm (refer back to Table 2.2), experienced a significantly larger increase in pupil size in response to semantic anomalies than did their less neurotic peers (cf. Fig. 4.2e). This effect suggests more cognitive load for highly neurotic individuals when a clash is encountered, potentially as a semantic anomaly is more "upsetting" for individuals that are generally more prone to be less calm than others.

In the interaction with the listener's political values, individuals with higher scores on the scale (i.e. more progressive listeners) were found to experience the largest increase in pupil sizes in response to a semantic anomaly (cf. Fig. 4.2f). This may seem incoherent at first, as it is generally conservative individuals that are considered to be less tolerant of ambiguity, uncertainty, and integrative complexity (Jost et al. 2003). However, it is possible for more progressive individuals to simply "engage" more with the anomaly in an attempt to understand the speaker (Winn et al. 2018), as progressive individuals are generally thought to be more tolerant of other perspectives. It is also conceivable that more progressive listeners engage in more anticipation as to what an upcoming segment might be, similarly to more empathetic individuals, or listeners in a good mood (Havas et al. 2007; Van Berkum et al. 2013; Van den Brink et al. 2010; Zadra and Clore 2011; again refer back to Sections 1.1.2.2 and 1.1.2.3). We will return to a broader discussion of this effect, and of the need for further research regarding the influence that an individual's political leaning has on language comprehension, in the General Discussion.



(a) Example interaction between time and item rating from the political values GAMM.

(b) Example interaction between time and item rating, visualizing the region where the effect is significant; from the political values GAMM.

(c) Visualization of the three-way interaction between time, item rating, and the listener's Openness, with pupil sizes on the z-axis (colour scale).

(d) Visualization of the three-way interaction between time, item rating, and the listener's Extraversion, with pupil sizes on the z-axis (colour scale).

(e) Visualization of the three-way interaction between time, item rating, and the listener's Neuroticism, with pupil sizes on the z-axis (colour scale).

(f) Visualization of the three-way interaction between time, item rating, and the listener's political values, with pupil sizes on the z-axis (colour scale).

Figure 4.2: Visualizations of the best GAM models for semantic anomalies.

4.5.3 Socio-Cultural Clashes

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	10.4359	4.6702	2.2346	0.0254
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	4.8820	5.4912	10.6321	< 0.0001
Item rating	8.9707	8.9993	156.0911	< 0.0001
Agreeableness	1.0021	1.0023	5.9364	0.0148
Time : rating	10.2859	12.3599	14.0620	< 0.0001
Agreeableness : time	1.0095	1.0111	4.8489	0.0273
Agreeableness : rating	15.6249	15.9703	50.9152	< 0.0001
Agr. : time : rating	52.7305	58.9510	12.7632	< 0.0001
Random structure				
Participant : time	209.8947	295.0000	31.7825	< 0.0001
Item	99.8421	101.0000	91.6456	< 0.0001

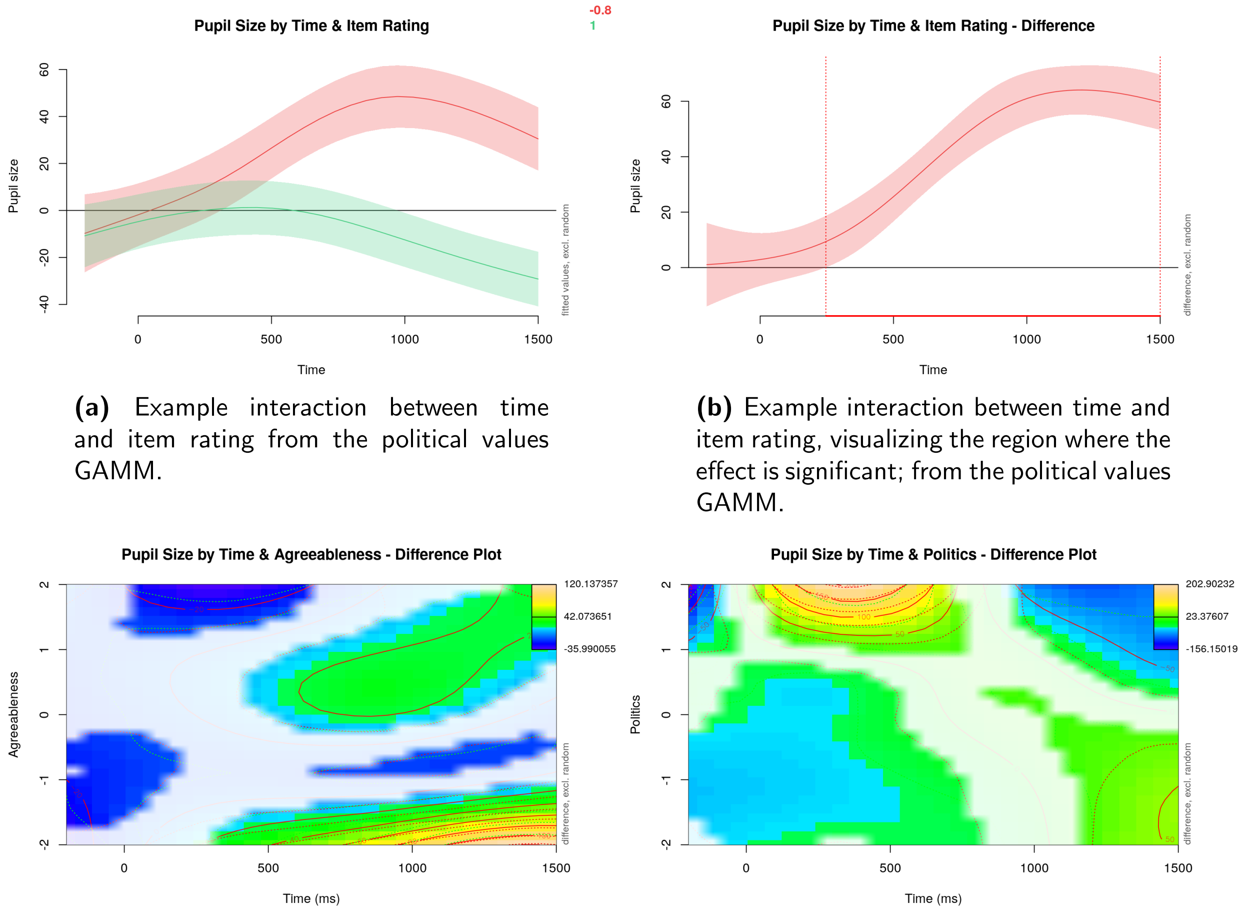
Table 4.6: Output of the GAMM for socio-cultural clashes that uses the listener's Agreeableness as the extra-linguistic predictor ($AIC = 3,275,652$).

In the modelling of pupil sizes in response to socio-cultural clashes, again a significant difference emerged between clashing and non-clashing items, starting at around 300ms after clash onset (cf. Figs. 4.3a and 4.3b). Two extra-linguistic variables were found to have a significant effect in a three-way interaction with time and item rating ($n = 252,915$ for both models; for model summaries, see Tables 4.6 and 4.7): Firstly, less agreeable individuals experienced a larger increase in pupil sizes than did their more agreeable peers when encountering a socio-cultural clash; cf. Fig. 4.3c). This is intuitively accessible, as less agreeable individuals are generally described as less cooperative, trustful, and sympathetic (refer back to Table 2.2). Secondly, increased pupil sizes were found for progressive-leaning individuals, i.e. those with higher scores on the political scale, when encountering a socio-cultural clash (cf. Fig. 4.3d). This effect is in line with the one found in response to semantic anomalies, discussed just above (cf. Fig. 4.2f); the only difference is that, in comparison, the effect for socio-cultural clashes appears to dissipate more quickly, and only affects listeners with the most progressive views on the political values scale.

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	9.5055	4.8923	1.9429	0.0520
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	4.7339	5.3261	9.8303	< 0.0001
Item rating	8.9640	8.9990	148.6886	< 0.0001
Political values	1.0029	1.0030	0.0014	0.9706
Time : rating	10.6050	12.5107	16.9430	< 0.0001
Political values : time	2.5902	2.7990	1.3550	0.2350
Political values : rating	15.6162	15.9561	61.8417	< 0.0001
Pol. : time : rating	57.4029	61.4898	15.6912	< 0.0001
Random structure				
Participant : time	208.5898	295.0000	33.4273	< 0.0001
Item	99.8501	101.0000	92.6467	< 0.0001

Table 4.7: Output of the GAMM for socio-cultural clashes that uses the listener's political values as the extra-linguistic predictor ($AIC = 3,275,300$).

This effect suggests that the processing of both semantic anomalies and socio-cultural clashes is influenced by the listener's political stance. Crucially, as no such effect was found in the modelling of pupil sizes in response to morpho-syntactic errors, this suggests that this intra-linguistic type of error may not "recruit" the listener's political stance, in contrast to the more semantic, world-knowledge related clashes.



(a) Example interaction between time and item rating from the political values GAMM.

(b) Example interaction between time and item rating, visualizing the region where the effect is significant; from the political values GAMM.

(c) Visualization of the three-way interaction between time, item rating, and the listener's Agreeableness with pupil sizes on the z-axis (colour scale).

(d) Visualization of the three-way interaction between time, item rating, and the listener's political values, with pupil sizes on the z-axis (colour scale).

Figure 4.3: Visualizations of the best GAM models for socio-cultural clashes.

4.6 Discussion

In this pupillometry experiment, we were able to identify several interactions between an extra-linguistic variable, average item rating, and time, specifically: The listener's Extraversion in the modelling of morpho-syntactic errors; the listener's Openness, Extraversion, Neuroticism, and political values, respectively, in pupillary responses to semantic anomalies; and the listener's Agreeableness and political values in the modelling of socio-cultural clashes. As expected, there was no one variable that influenced the comprehension of all three clash types across the board. The only predictors that influenced changes in pupil size in response to more than one type of clash were the listener's Extraversion and political values. In the case of Extraversion, it was more introverted listeners who experienced a larger change in pupil size, i.e. a larger cognitive load, when encountering a morpho-syntactic error or semantic anomaly; conversely, a high score on the political scale (suggesting a more progressive outlook) was associated with a larger change in pupil size in response to semantic violations and socio-cultural clashes. This pattern highlights the interesting "in-between" status of semantic anomalies: While morpho-syntactic errors violate intra-linguistic rules and do not rely on world knowledge, and while socio-cultural clashes derive all their "strangeness" from what the listener infers about the speaker, and from how this inference clashes with the world as they know it, semantic violations are situated in between the two. As semantic mismatches between the verb and its object, they are not quite pure intra-linguistic violations, but at the same time they also do not require "access" to the speaker's identity to be anomalous; in the stimuli used in this dissertation, they are anomalous independently of who utters the sentence that contains them (unless embedded in a context that explicitly renders them non-anomalous; recall the animate peanuts in Nieuwland and Van Berkum 2006).

As discussed previously, the observed effects of the listener's political values suggest that progressive individuals may "engage" more with the semantic anomaly and the socio-cultural clash, rather than to avoid it. Recall firstly that increased pupil size was found to measure intentional attentional engagement, as per Winn et al. (2018); and secondly that, in the ratings study, it was progressive listeners who seemed to have a wider "range" on which to rate items, such that the difference in ratings between non-anomalous vs. anomalous stimuli was a lot wider than for conservative listeners (cf. Section 2.5.2, and Fig. 2.3c in particular). At the same time, the finding could also suggest more anticipation based on speaker clues;

more research into what factors into an individual's political stance, and how it interacts specifically with language comprehension, is needed. We will return to this very general issue in the General Discussion chapter below.

A significant interaction with the listener's Extraversion was observed in the models for morpho-syntactic errors and semantic anomalies. As was already discussed above, the effect for semantic anomalies reflects the finding from the item ratings study: Introverted individuals seemed to experience more cognitive effort when encountering an anomaly (cf. Fig. 4.2d), and also rated sentences containing such an anomaly worse than their extraverted counterparts (cf. Fig. 2.3b). At the same time, the relationship between Extraversion and language comprehension is not as clear when it comes to the processing of morpho-syntactic errors. Much like for semantic anomalies, introverted individuals seem to experience greater cognitive load when processing a morpho-syntactic error (cf. Fig. 4.1c) – but they then rated the sentence containing the error *better* than their extraverted peers (cf. Fig. 2.1e). It is not immediately clear where this disconnect stems from – if introverted individuals simply were exposed less to unusual stimuli due to fewer social encounters, or fewer *diverse* social encounters, they would be expected to experience greater cognitive load when encountering an unusual utterance, and at the same time to rate the utterance as less acceptable than their extraverted peers. This seems to be the case for semantic anomalies, but not for morpho-syntactic errors; we will return to how the type of clash may interact with the listener's Extraversion in the General Discussion below.

Furthermore, clashes triggered significantly larger pupil sizes in the same time frame for all clash types (cf. Figs. 4.1b, 4.2b, and 4.3b), and individual difference effect did not surface later in the modelling of socio-cultural clashes as compared to the other two types (compare the surface plots in Figs. 4.1, 4.2, and 4.3). This suggests that these listener-internal factors are considered in language comprehension rather early, at the same stage as syntactic information, rather than being integrated in a second step at a later time (see also Hagoort et al. 2004; Knoeferle et al. 2005; Nieuwland and Van Berkum 2006; Van Berkum et al. 2005, and recall our earlier discussion of one- and two-step models in Chapter 1).

We will now move on to the second pupillometry experiment, which uses the same general paradigm as the pupillometry study in this current chapter. However, it differs in that it firstly introduces the Disgust Sensitivity variable, and in that secondly, participant recruitment was expanded beyond the undergraduate student pool. After this, results from all four experiments will be discussed in the General Discussion in Chapter 6.

Chapter 5

Experiment IV: Pupillometry & Disgust[1]

This second pupillometry experiment largely addresses the same questions as Chapter 4, and thus used the same experiment procedure as Experiment III. However, it differs from the previous pupillometry experiment in that it introduced the listener's Disgust Sensitivity as an additional listener-related variable. Disgust Sensitivity has been found to be strongly linked to feelings of morality, purity, and – crucially for this dissertation – political orientation and outgroup stigmatization (for a detailed discussion on Disgust Sensitivity, and on how it relates to political views and stigmatization, see Section 1.1.2.4). It has, to the best of our knowledge, not been investigated previously with regards to language comprehension. We predicted Disgust Sensitivity to significantly affect changes in pupil size, particularly for socio-cultural clashes. Specifically, we predicted that individuals more sensitive to disgusting stimuli, who thus can be assumed to have a stronger desire towards pathogen avoidance and can thus be assumed to engage in more outgroup stigmatization, would experience a larger cognitive load upon encountering a statement that clashes with established gender stereotypes.

In addition, this present study used a different participant sampling strategy than the previous three experiments in this dissertation: Participants were recruited not only from the undergraduate linguistics participant pool, but also externally, allowing for a wider range of ages and "experiences in the world," with the goal to analyze a sample that better represented the general population than the common undergraduate student convenience sample.

[1] Parts of this chapter were presented at the *Annual Meeting of the Cognitive Science Society* (Montréal, QC, 25-27 July 2019), with subsequent publication in the conference proceedings.

5.1 Participants

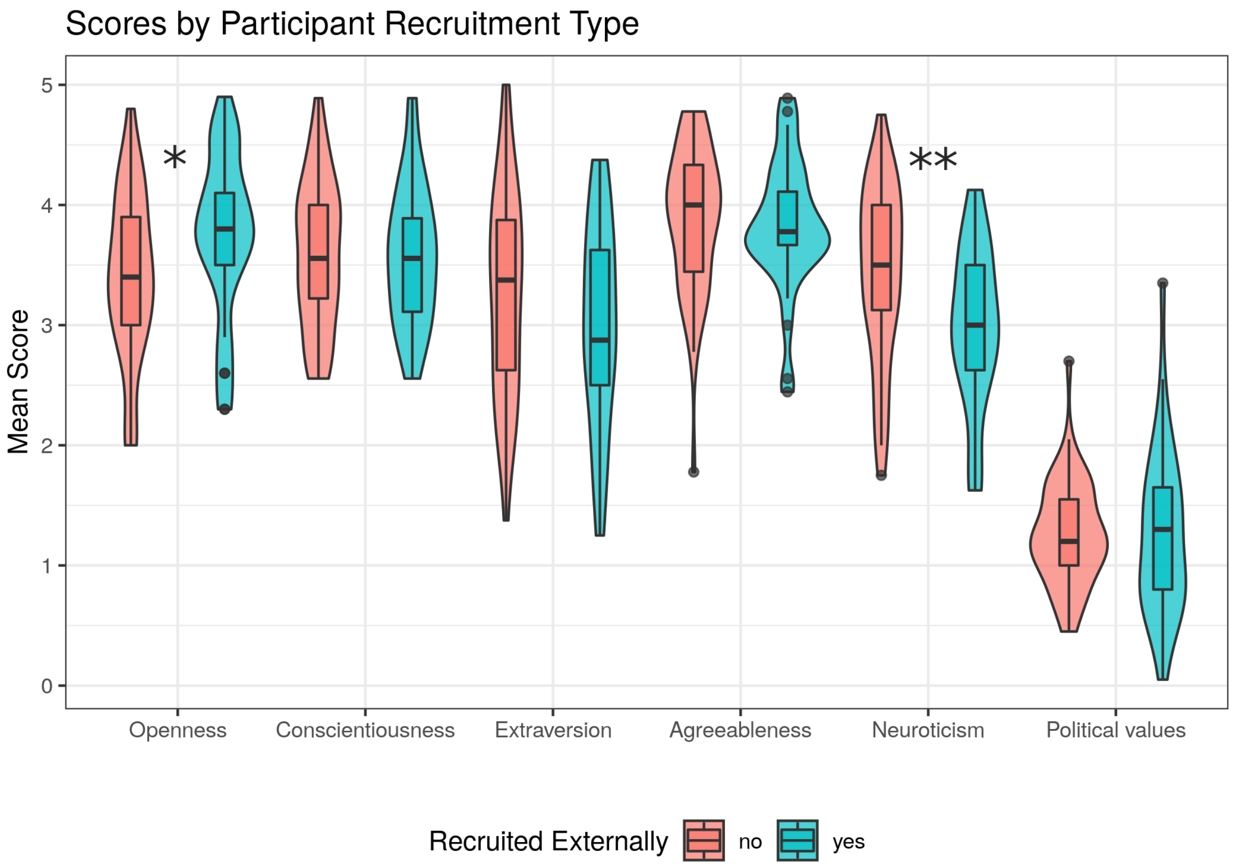


Figure 5.1: Visualization of Big Five and political values score distributions between the two different participant recruitment strategies.

82 participants in total completed this experiment. 49 (60%) were recruited from the university's undergraduate linguistics pool, and received course credit for their participation. Another 33 (40%) were recruited from the general population, not limited to the University of Alberta campus or to an academic background, and received a small monetary compensation for their participation. A comparison of average Big Five and political values scores between the two participant groups is visualized in Fig. 5.1. Note that significant differences between the participant groups were only found for the Neuroticism sub-scale, where externally recruited participants were found to be significantly less neurotic than those participants recruited from the undergraduate linguistics pool ($mean_{external} = 2.98, SD_{external} = 0.67; mean_{internal} = 3.43, SD_{internal} = 0.73; t(72.802) = -2.86, p < 0.01$), and for the Openness sub-scale, where externally recruited participants were found to be significantly more

open ($mean_{external} = 3.75, SD_{external} = 0.68; mean_{internal} = 3.41, SD_{internal} = 0.69; t(69.303) = 2.18, p = 0.03$). These differences may stem from different motivations for participants to sign up for the study: Whereas undergraduate pool students are required to obtain a certain amount of course credit via research participation (not necessarily participation in this particular study, but *a* study or a set of studies), external participants had no obligation to do so, and participated entirely from their own volition. It is possible that highly neurotic individuals proportionally sign up less for experimental research than their less neurotic peers, and that more open individuals are more amenable to the idea, thus resulting in a different trait distribution in the somewhat self-selected sample.

Data from eight participants was removed as their comprehension question accuracy rates were below 80% ($min = 75\%, max = 100\%, mean = 93.7\%, median = 96.4\%, SD = 6.6\%$), and comprehension or attention to the experiment could hence not be guaranteed; or as information given on the language background questionnaire precluded their data from inclusion in the analyses. Data from 728 trials (roughly 8% of trials) was removed due to issues during recording that resulted in more than 33% of sampling points on a given trial being recorded as `N/A`. Thus, analyses in this chapter are based on the data from 74 participants (males/females = 16/58; native/non-native speakers of English = 60/14; age = 17–83; mean [SD] = 25 [12.7] years).

5.2 Materials

For consistency, the same stimuli from Experiments I through III were re-used, and presented in the same format as in Experiments I and III (see Table 2.1 for the template, and Section A in the Appendix for the full list of items; for details, refer back to Section 2.2 and 4.2).

5.3 Procedure

The procedure of the main pupillometry experiment was, for the sake of comparability, the same as in Experiment III; for details, refer back to Section 4.3.

5.4 Post-Tests

The post-tests presented to the participants were largely the same as in the previous three experiments, as they included the same Big Five personality assessment and the same language background questionnaire as in Experiments I through III, and an assessment of political views. To assess the influence of Disgust Sensitivity on language comprehension, the *Disgust Scale - Revised* (*DS-R*; Haidt, McCauley & Rozin, 1994, modified by Olatunji et al. 2007), which was also used in, for example, Ahn et al. (2014) and Inbar et al. (2009, 2011), was administered to participants; the full scale can be found in Appendix B.6. Additionally, the previous political questionnaire was replaced with a *Wilson-Patterson*-type test (Wilson and Patterson 1968) – a slightly shorter, more established questionnaire compared to the test used in the first three experiments, the full version of which can be found in Appendix B.4. The test was also chosen for results to be more directly comparable to recent research involving political values and Disgust Sensitivity (Ahn et al. 2014; Hatemi and Verhulst 2015; Jost et al. 2003; Smith et al. 2011). Note that the Wilson-Patterson scale is a conservativism scale; as such, high scores signify a conservative outlook, as opposed to the political questionnaire used in the previous three experiments, where a high score signified a progressive outlook.

5.5 Results

The raw pupillometry data was pre-processed and modelled in the same way as in Experiment III (refer back to Section 5.5 for details), with the additional post-test scores being added as additional numeric variables. Note that again all numerical predictors were scaled and centered to avoid unintended effects of different orders of magnitude between predictors. Prior research has reported systematically higher Disgust Sensitivity among women as compared to men (Al-Shawaf et al. 2018; Sparks et al. 2018); in this present study, only a non-significant tendency in the same direction was found in a two-sample t-test ($mean_{male} = 1.78, SD_{male} = 0.68; mean_{female} = 2.06, SD_{female} = 0.58; t(28.678) = -1.62, p = 0.12$). Note that interactions with speaker and listener gender, or with native speaker status, were not found to be significant in any of the models reported below.

5.5.1 Morpho-Syntactic Errors

In the modelling of pupil sizes in response to morpho-syntactic errors, **Neuroticism** and **Extraversion** were found to be significant extra-linguistic predictors, as observed in a three-way interaction with time and item rating in the respective GAM models (pupil size samples $n = 400,658$; see Tables 5.1 and 5.2 for model summaries).

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	18.0742	4.4884	4.0269	0.0001
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	6.1963	6.8692	15.0198	< 0.0001
Item rating	8.9104	8.9954	151.1357	< 0.0001
Neuroticism	1.0029	1.0030	0.0362	0.8504
Time : rating	13.1014	14.9135	28.6975	< 0.0001
Neuroticism : time	3.4956	3.5810	1.6163	0.1279
Neuroticism : rating	15.7789	15.9909	71.0375	< 0.0001
Neur. : time : rating	54.5596	60.4854	10.6515	< 0.0001
Random structure				
Participant : time	492.8862	664.0000	34.4956	< 0.0001
Item	100.8714	102.0000	89.4889	< 0.0001

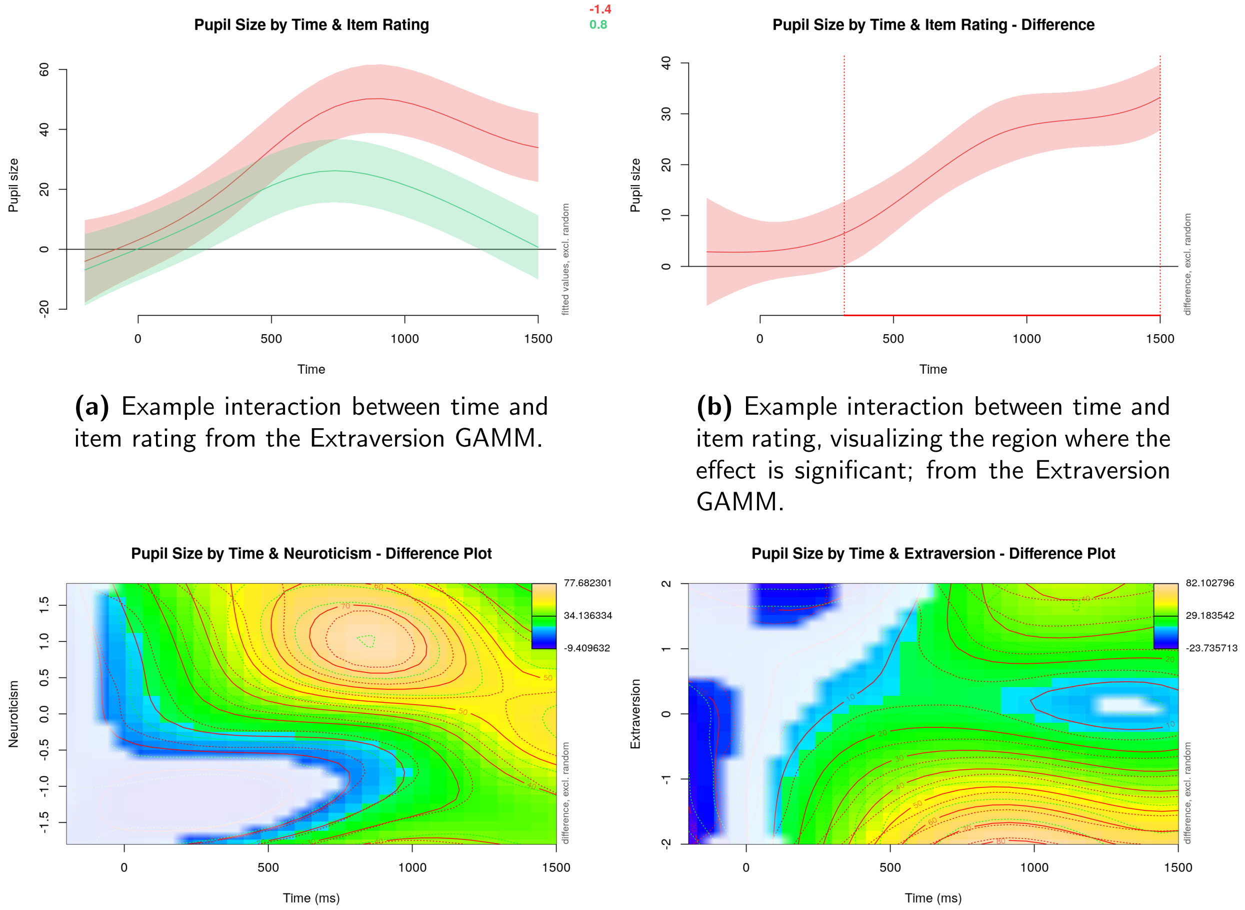
Table 5.1: Output of the GAMM for morpho-syntactic errors that uses the listener's Neuroticism as the extra-linguistic predictor ($AIC = 5,244,179$).

In both models, a significant interaction between item condition (i.e. average item rating) and time was found: Pupil sizes increased significantly around 350ms after listeners encountered a morpho-syntactic error (cf. Figs. 5.2a and 5.2b). In the two significant three-way interactions, it was more neurotic (cf. Fig. 5.2c) and less extraverted individuals (cf. Fig. 5.2d) that experienced an increase in pupil size when encountering a morpho-syntactic error, with the Extraversion effect replicating the effect found in the modelling of morpho-syntactic errors and semantic anomalies in Experiment III (cf. Figs. 4.1c and 4.2d). As discussed previously, this could be a result of more introverted individuals simply being exposed less to unusual stimuli due to fewer, or less diverse, social interactions; we will return to a broader discussion of the Extraversion effect, and whether it influences cognitive load during the experiment directly, or via decreased prior exposure to unusual stimuli, in the General Discussion chapter below.

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	17.5706	4.4512	3.9474	0.0001
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	6.3068	6.9739	15.8327	< 0.0001
Item rating	8.9177	8.9959	167.9871	< 0.0001
Extraversion	1.0080	1.0084	0.0173	0.8997
Time : rating	13.3430	15.0046	32.1432	< 0.0001
Extraversion : time	3.3753	3.4725	0.8638	0.2578
Extraversion : rating	15.8564	15.9950	81.2962	< 0.0001
Extr. : time : rating	54.4306	60.2798	9.9324	< 0.0001
Random structure				
Participant : time	495.0240	664.0000	35.8167	< 0.0001
Item	100.8648	102.0000	89.0313	< 0.0001

Table 5.2: Output of the GAMM for morpho-syntactic errors that uses the listener's Extraversion as the extra-linguistic predictor ($AIC = 5,244,044$).

While no effect of Neuroticism on pupil sizes was found in Experiment III, where an effect was found for Extraversion only (refer back to Section 4.5.1), the listener's Neuroticism scores affected item ratings in the first study in this dissertation. Recall that Neuroticism was one of the two Big Five sub-scales that was found to differ significantly between the two participant recruitment groups in this experiment (cf. Fig. 5.1), but that, at the same time, the overall Neuroticism distribution did not differ significantly between the participant samples of the two pupillometry studies (refer back to Table 1.1 and Fig. 1.1 in the Introduction). Interestingly, the effect reported for the listener's Neuroticism scores in this pupillometry study runs in the "opposite" direction compared to the effect observed in the ratings study: As visualized in Fig. 2.1f, highly neurotic listeners rated non-anomalous items worse, but erroneous items *better* than their less neurotic counterparts. This result does not align immediately with the influence that Neuroticism was found to have in this present pupillometry study, where it was highly neurotic individuals that showed a *larger* increase in pupil size in response to errors, suggesting more cognitive effort. This suggests that subtle differences in the participant samples, even if the samples look similar to each other in terms of average trait scores, may be reflected in the results; or, alternatively, this finding may highlight differences in what exactly is measured by the two very different experimental paradigms – off-line ratings using a conscious mouse-click action after listeners have processed the entire



(a) Example interaction between time and item rating from the Extraversion GAMM.

(b) Example interaction between time and item rating, visualizing the region where the effect is significant; from the Extraversion GAMM.

(c) Visualization of the three-way interaction between time, item rating, and the listener's Neuroticism, with pupil sizes on the z-axis (colour scale).

(d) Visualization of the three-way interaction between time, item rating, and the listener's Extraversion, with pupil sizes on the z-axis (colour scale).

Figure 5.2: Visualizations of the best GAM models for morpho-syntactic errors.

stimulus, vs. a continuous assessment of processing load as the item unfolds. We will return to a broader discussion of what the different paradigms investigate – conscious ratings after the fact vs. automated comprehension as it happens – in the General Discussion chapter.

5.5.2 Semantic Anomalies

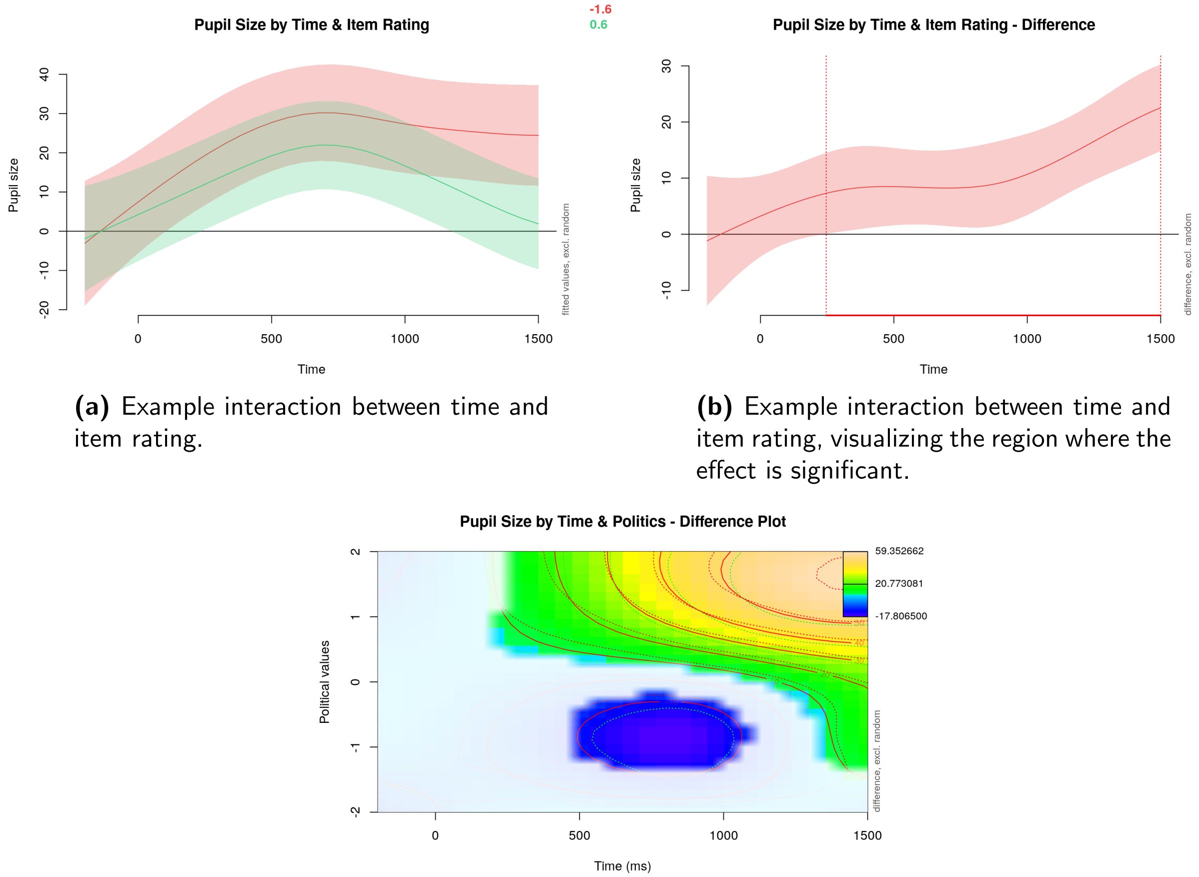
Again, in the modelling of semantic anomalies, anomalous utterances were found to elicit significantly larger pupil sizes than non-anomalous ones, starting approximately 250ms after anomaly onset (cf. Figs. 5.3a and 5.3b). Of special interest for the research questions in this dissertation was that only the listener's **political values** were found to interact significantly with item rating and time ($n = 359,050$; for the full model summary, see Table 5.3). Specifically, it was individuals with higher scores on the political values scale – i.e. more conservative listeners[2] – that experienced a larger cognitive load when encountering a semantic anomaly (cf. Fig. 5.3c).

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	15.8662	4.3184	3.6741	0.0002
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	5.6948	6.3724	15.0944	< 0.0001
Item rating	8.8134	8.9819	156.7196	< 0.0001
Political values	2.3699	2.3948	2.1141	0.1322
Time : rating	14.5393	15.6844	31.0968	< 0.0001
Political values : time	4.0893	4.3338	1.3794	0.2026
Political values : rating	14.9388	15.7410	35.8636	< 0.0001
Pol. : time : rating	46.0817	53.8418	9.3957	< 0.0001
Random structure				
Participant : time	481.8050	664.0000	31.2707	< 0.0001
Item	100.4849	102.0000	76.3538	< 0.0001

Table 5.3: Output of the GAMM for semantic anomalies that uses the listener's political values as the extra-linguistic predictor ($AIC = 4,687,172$)

While an effect was observed for semantic anomalies in the first pupillometry study (cf. Section 4.5.2, and specifically Fig. 4.2f), it runs in the opposite direction – in the previous experiment, it was *progressive* listeners who experienced a larger increase in pupil size. However, note the stark difference in timing: Whereas in Experiment III, progressive listeners experienced a significantly larger pupil size as early as 150ms after anomaly onset, this significant effect only surfaced at around 400ms for conservative listeners in this current

[2] Recall that the Wilson-Patterson scale used in this final experiment has an opposite polarity to the political values test used in the other three experiments.



(a) Example interaction between time and item rating.

(b) Example interaction between time and item rating, visualizing the region where the effect is significant.

(c) Interaction between time, item rating, and political values.

Figure 5.3: Visualizations of the political values GAM model for semantic anomalies.

experiment. We will return to a discussion of these two opposite effects, and the difference in timing between them that may signify an involuntary "gut reaction" even in individuals that take no issue with the actual statement, in the General Discussion.

Further, three extra-linguistic variables in addition to the listener's political values were found to be significant in the previous experiment, namely Openness, Extraversion, and Neuroticism. Recall that two of these three variables, Openness and Neuroticism, were found to differ significantly between the two recruitment groups in the current study (cf. Fig. 5.1), where externally recruited participants were more open and less neurotic than their undergraduate student peers. One may be inclined to link the absence in effects to the significant difference in scores between the two different recruitment groups; however, even if externally recruited participants are less neurotic and more open than their student counterparts, comparing the two participant samples of Experiment III and IV, it becomes obvious that the two samples *overall* do not significantly differ from each other in their Neuroticism and Openness scores (refer back to Table 1.1 and Fig. 1.1 in the Introduction). This suggests that there may be other differences between the two participant samples (or, more generally, between *any* two participant samples) that cannot be measured by this current set of statistical models. We will return to this issue, and the more general question of which variables to control for (and where to stop), in the General Discussion below.

5.5.3 Socio-Cultural Clashes

In the modelling of pupil sizes in response to socio-cultural clashes, clashing items elicited a significantly larger pupil size compared to non-clashing items, starting around 400ms after clash onset (cf. Figs. 5.4a and 5.4b). As predicted, the listener's Disgust Sensitivity was found to be a significant listener-internal variable, in addition to Extraversion and Openness ($n = 496,145$ for all models; for the full model summaries, see Tables 5.4, 5.5, and 5.6): Listeners with higher Disgust Sensitivity showed a larger increase in pupil size when they encountered a socio-cultural clash than did listeners less sensitive to disgust (cf. Fig. 5.4e). This can be traced back to the strong relation between pathogen avoidance and outgroup stigmatization (refer back to Section 1.1.2.4). Note that, as per the combination test discussed in Section 4.5, the effects of Disgust Sensitivity remained – it only became slightly less pronounced – when combining it with Openness and Extraversion, i.e. the two Big Five traits that were found to be significant predictors in modelling pupil size in response to socio-cultural clashes in this experiment. To the best of our knowledge, this result shows for the first time that Disgust Sensitivity affects language comprehension right as it happens.

In the interaction with Openness, it was less open individuals that experienced a larger cognitive load when encountering a socio-cultural clash (cf. Fig. 5.4c). It is interesting to note that this effect was not observed in response to socio-cultural clashes in the previous pupillometry experiment, which may again be related to the difference in participant recruitment already discussed for semantic anomalies just previously. The Openness effect in this experiment is in line with the effect of Openness on item ratings (cf. Section 2.5.3 and Fig. 2.4c), where less open individuals rated clashing items much worse than their more open peers. We will return to this effect, and specifically how it runs opposite to the effect that the listener's Openness seems to have on the processing of morpho-syntactic errors and semantic anomalies, in the General Discussion; the difference may very well be related to

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	14.2651	4.1288	3.4550	0.0006
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	5.9729	6.6562	18.7944	< 0.0001
Item rating	8.7248	8.9728	126.0199	< 0.0001
Disgust Sensitivity	1.0009	1.0009	9.8377	0.0017
Time : rating	13.7975	15.2517	26.9851	< 0.0001
Disgust Sensitivity : time	2.5235	2.6540	4.4291	0.0244
Disgust Sensitivity : rating	15.6830	15.9774	46.1007	< 0.0001
Disgust : time : rating	47.3150	54.1664	7.6567	< 0.0001
Random structure				
Participant : time	488.3972	664.0000	33.3408	< 0.0001
Item	101.0948	102.0000	127.3553	< 0.0001

Table 5.4: Output of the GAMM for socio-cultural clashes that uses the listener's Disgust Sensitivity as the extra-linguistic predictor ($AIC = 6,479,425$).

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	14.9019	4.1844	3.5613	0.0004
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	5.9502	6.6300	18.5532	< 0.0001
Item rating	8.6886	8.9659	116.8285	< 0.0001
Openness	1.0037	1.0038	0.2816	0.5964
Time : rating	14.3946	15.5625	21.8721	< 0.0001
Openness : time	1.0136	1.0155	0.1463	0.7063
Openness : rating	15.6997	15.9716	86.1973	< 0.0001
Openn. : time : rating	53.6238	59.2077	12.5140	< 0.0001
Random structure				
Participant : time	491.8005	664.0000	32.6620	< 0.0001
Item	101.0934	102.0000	126.0124	< 0.0001

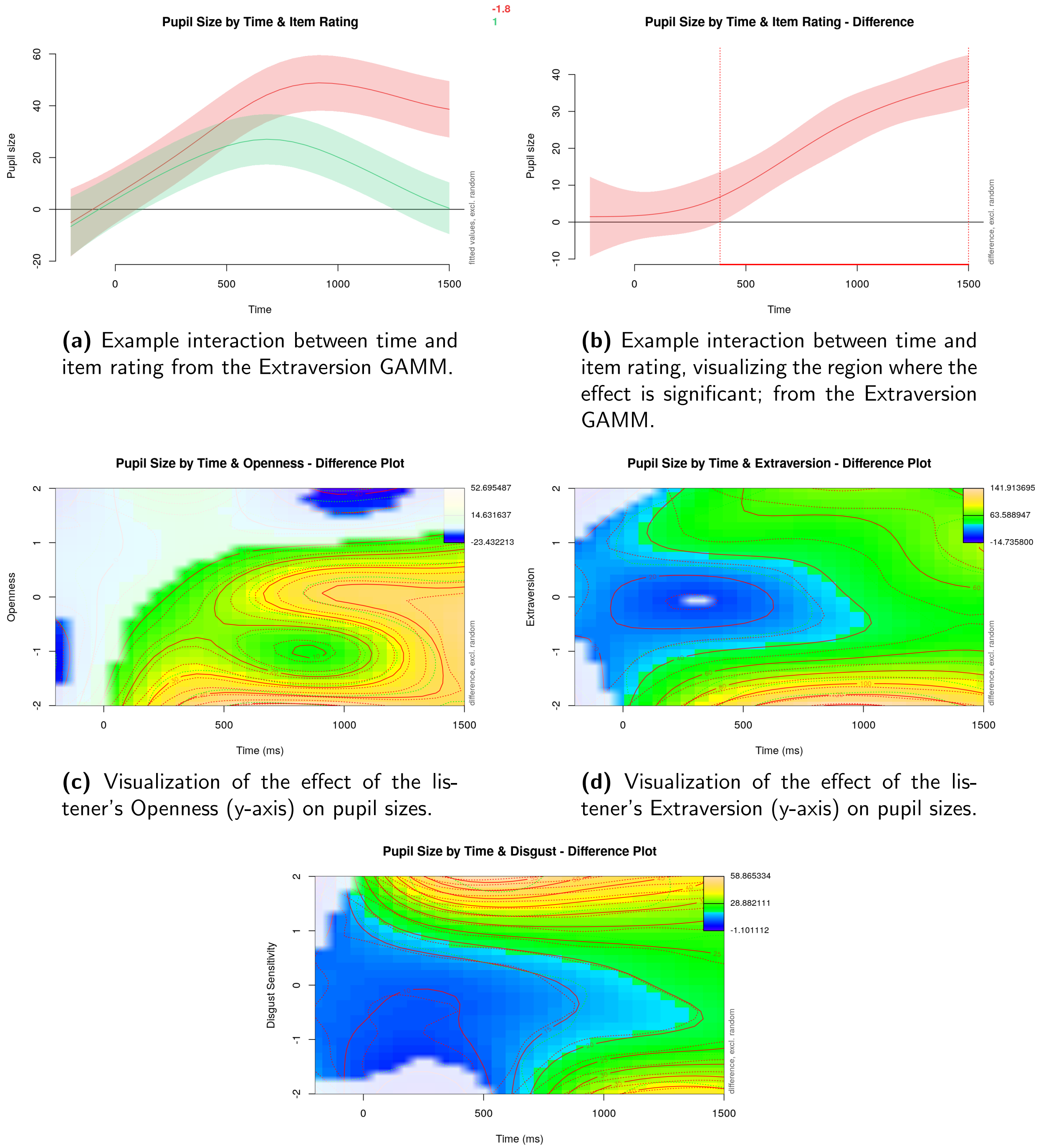
Table 5.5: Output of the GAMM for socio-cultural clashes that uses the listener's Openness as the extra-linguistic predictor ($AIC = 6,478,474$).

Parametric coefficients	*Estimate*	*Std. Error*	*t-score*	*p-value*
(Intercept)	14.5440	4.1778	3.4812	0.0005
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Time	6.0012	6.6847	18.8488	< 0.0001
Item rating	8.6887	8.9660	130.1096	< 0.0001
Extraversion	1.0294	1.0308	0.3058	0.5748
Time : rating	14.1766	15.4686	26.1892	< 0.0001
Extraversion : time	1.2734	1.2885	0.1382	0.6802
Extraversion : rating	15.7732	15.9890	78.9764	< 0.0001
Extr. : time : rating	57.6740	62.1340	10.8406	< 0.0001
Random structure				
Participant : time	490.2904	664.0000	33.5104	< 0.0001
Item	101.0817	102.0000	126.5691	< 0.0001

Table 5.6: Output of the GAMM for socio-cultural clashes that uses the listener's Extraversion as the extra-linguistic predictor ($AIC = 6,478,644$).the

the “nature” of the clash: The processing of clashes directly involving the inferred identity of the speaker may be influenced differently by the listener's Openness than clashes that do not rely on this particular feature.

Further, more introverted individuals seemed to experience a larger cognitive load when encountering a socio-cultural clash (cf. Fig. 5.4d). Much like the effect reported for Openness just above, this effect was also not observed in the first pupillometry study. However, Extraversion – unlike Openness – was not among the variables that was found to differ significantly between the two recruitment strategies, so that an explanation along the lines of participant recruitment is not applicable here. At the same time, this Extraversion effect is highly reminiscent of the same effect observed in response to morpho-syntactic errors in both pupillometry studies (cf. Figs. 4.1c and 5.2d). As mentioned previously, we will return to the “opposing”’ effects of Extraversion in the General Discussion below.



(a) Example interaction between time and item rating from the Extraversion GAMM.

(b) Example interaction between time and item rating, visualizing the region where the effect is significant; from the Extraversion GAMM.

(c) Visualization of the effect of the listener's Openness (y-axis) on pupil sizes.

(d) Visualization of the effect of the listener's Extraversion (y-axis) on pupil sizes.

(e) Visualization of the effect of the listener's Disgust Sensitivity (y-axis) on pupil sizes.

Figure 5.4: Visualizations of the best GAM models for socio-cultural clashes.

5.6 Discussion

Overall, as in the other three experiments, this present pupillometry study showed that listener-internal variables influence language comprehension; and that, as already discussed for Experiment III previously, they do so right as comprehension happens. That is, listeners experience a larger cognitive load, as indicated through a significantly larger increase in pupil size, when they come across an unusual segment, and this increase in pupil size is modulated by certain listener-internal factors. In regards to timing, it is important to note that we again, just as in Experiment III, did not see systematic differences between socio-cultural clashes on the one hand, and morpho-syntactic errors and semantic anomalies on the other (refer back to Figs. 5.2b, 5.3b, and 5.4b). Individual difference effects again did not surface later in the modelling of socio-cultural clashes as compared to the other two types (compare the surface plots in Figs. 5.2, 5.3, and 5.4) This lends further support to these listener-internal factors being considered early, in the same step as syntactic information (Hagoort et al. 2004; Knoeferle et al. 2005; Nieuwland and Van Berkum 2006; Van Berkum et al. 2005, and also refer back to Chapter 1).

The most crucial finding in this experiment is that Disgust Sensitivity, which was introduced as a new listener-related variable, indeed was found to be a significant listener-internal predictor in the modelling of pupil sizes in response to socio-cultural clashes. That is, Disgust Sensitivity seems to modulate the cognitive load that listeners experience when they encounter a statement clashing with established gender stereotypes: Someone who is highly sensitive to disgusting stimuli seems to experience a larger cognitive load when encountering a socio-cultural clash. This suggests that the link between Disgust Sensitivity and out-group stigmatization is reflected in automated language comprehension (also refer back to Section 1.1.2.4). We will discuss this finding in the broader context of the Behavioural Immune System in the General Discussion chapter.

In terms of significant predictors, there is only little overlap between this pupillometry study and the ratings experiment,[3] namely an interaction involving Openness in response to socio-cultural clashes: Less open individuals experienced a larger pupil size when coming across items clashing with established gender stereotypes, and also rated these items sig-

[3] An overview of significant extra-linguistic effects across the four experiments can be found in Table 6.1 in the General Discussion chapter.

nificantly worse than their more open peers (cf. Figs. 5.4c and 2.4c). However, between the ratings experiment and this current pupillometry study, three effects ran into an opposite direction – specifically, the effects of Neuroticism and Extraversion on the processing of morpho-syntactic errors, and the effect of political values on the processing of semantic anomalies. While the overall Big Five trait distribution within the participant sample in this fourth experiment does not seem out of line compared to values reported in the literature (see Table 1.1 and Fig. 1.1), one explanation for this phenomenon could be the different sampling method used in this final experiment. As mentioned previously, the participant sample was recruited both externally and from the undergraduate linguistics pool, which resulted in markedly different distributions regarding age, and the Neuroticism and Openness sub-scales of the Big Five test (cf. Fig. 5.1). At the same time, the overall distributions between the participant samples in the different studies are not significantly different. As discussed in an earlier section in this chapter, this could suggest that the participant samples may differ in aspects not currently controlled for that affect results. We will return to this, and how this issue could be amended and controlled for more tightly in future research, in the General Discussion.

An explanation for the Extraversion effect on morpho-syntactic errors, where introverted participants in both pupillometry experiments experienced a larger increase in pupil size compared to their extraverted peers, but where it was extraverted individuals that rated erroneous items worse than their introverted peers, may lie in the vastly different paradigms: As already mentioned in Section 4.5.1, pupillometry assesses language comprehension on-the-fly, without any conscious input being required; the ratings experiment, on the other hand, assesses the comprehension of an utterance after the fact, with the participant having to engage in a conscious mouse-click action to complete the task. We will return to a discussion of how the Extraversion trait may interact with these different paradigms in the General Discussion below.

By far the “trickiest” effect, within the context of the previous three experiments, is that of the listener’s political values on the processing of semantic anomalies. In the earlier experiments, it was progressive listeners who rated anomalous items worse than their more conservative peers, and who also experienced a larger pupil size in response to those anomalous items. In this current experiment, the opposite result surfaced – namely, it was *conservative* listeners who experienced larger cognitive load when encountering a semantic anomaly. This could suggest one of two things: Firstly, it is possible that the participant sample in this current study, where recruitment was expanded to include non-academic participants, differs significantly from the samples in the other two experiments, *beyond* what can be captured in a single score assessing political outlook (recall that the externally recruited sample did not differ significantly in their political values from the undergraduate sample). We will return to this very general caveat, and potential ways in which it could be addressed, in the General Discussion. A second explanation for this opposite effect may be rooted in the "subtlety" of effects detected using the pupillometry paradigm, and the differences in timing that were observed. Whereas the effect for progressive listeners in Experiment III surfaced very early, around 150ms after clash onset, the effect found for conservative listeners in this current study was significant much later, at around 400ms. It is possible that, even if participant samples are similar to each other in terms of their personality and political values distribution, simply because they were run at different times on different days, or as they may have read a political news piece that they really enjoyed, or that clashed with their world views and made them upset, a pupillometry experiment may detect differing effects. That is to say, very subtle differences – even how the experimenter greeted them the morning of the experiment session (see also unintentional biosocial and psychosocial experimenter effects; e.g. Chapman et al. 2018; Rosenthal 1976; Sheldrake 1998) – may change the participant's outlook sufficiently, in a temporary fashion, to trigger "opposing" pupillometry results. We will return to this very broad issue, and ideas on how to navigate the small space between controlling for potentially relevant variables and yet not overloading either the participant or the statistical model, in the General Discussion just below.

Chapter 6

General Discussion

Results from all four experiments in this dissertation indicate that, as expected, several variables related to the internal state of the listener, and related to information inferred about the speaker, influence spoken language comprehension. Across the three different experimental paradigms, several Big Five traits, political values, and Disgust Sensitivity (in Experiment IV) were found to influence the comprehension of three different types of clashes. An overview of effects is given in Table 6.1, with the different clash types and extra-linguistic variables presented on the y-axis, and the four experiments on the x-axis. Considering that extra-linguistic effects were found to surface in ERP signatures at around 250ms (Van Berkum et al. 2009), 300ms (Hagoort et al. 2004), or 400ms (Nieuwland and Van Berkum 2006; Van Berkum et al. 2005), and that pupil size is a much more slow-moving measure compared to ERP's (Rij et al. 2019), the significant effects of extra-linguistic variables around 250ms or 400ms in the pupillometry studies of this dissertation support the notion that extra-linguistic variables are considered in the same step as syntactic information.[1] Additionally, effects of linguistic form, or syntactic agreement, were found to not appear earlier than effects of semantic or socio-cultural mismatch (compare, for example, Figs. 5.2d and 5.4d), further supporting an account in which various different types of information, syntactic *and* extra-linguistic, influence language comprehension in one single, early step.

We will now discuss the different types of variables that were found to significantly affect language comprehension in the four experiments in more detail below.

[1] For details, refer back to the discussion of one- and two-step models, and how context is considered early in language comprehension, in Section 1.1 of the Introduction.

	Experiment			
	I	**II**	**III**	**IV**
	Ratings	**SPL**	**Pupil**	**Pupil**
Morpho-syntactic errors				
Open.	more open			
Consc.		less consc.		
Extr.	more extr.		less extr.	less extr.
Neur.	less neur.			more neur.
Pol.	progr.			
Semantic anomalies				
Open.			more open	
Extr.	less extr.		less extr.	
Agr.	less agr.			
Neur.	gender int.		more neur.	
Pol.	progr.		progr.	cons.
Socio-cultural clashes				
Open.	less open	gender int.		less open
Extr.		gender int.		less extr.
Agr.			less agr.	
Pol.	progr.		progr.	
Disg.	N/A	N/A	N/A	more disg.

Table 6.1: A summary of significant extra-linguistic effects found across all four experiments. Cells indicate which listeners **rated clashing items worse** (in Experiment I); experienced a **greater delay in response times** (Experiment II); or experienced a **larger change in pupil size** (Experiments III and IV) in response to clashing items. Note that, in Experiment II, significant interactions are more complex than for the other three experiments and are difficult to capture in a table; only interactions with item rating are indicated here. For details on these and additional interaction effects, refer back to Section 3.5.

6.1 Listener-Internal Variables

As predicted, morpho-syntactic errors, semantic anomalies, and socio-cultural clashes were affected differently by listener-internal variables. There was no one variable that affected comprehension across the board, either across (1) all three clash types investigated, or (2) across the three different experimental methods used (cf. Table 6.1). This suggests that the comprehension of the three different clash types recruits different kinds of information; that is, varying aspects of the listener's personality modulate the comprehension of the distinct types of clashes in different ways. For example, the listener's Conscientiousness only surfaced as a significant predictor of button-press times in response to morpho-syntactic errors (cf. Fig. 3.1a), but not in response to any other clash type, and in any of the other three experiments. The traits of Extraversion, political values, and Openness, on the other hand, influenced responses across a range of clash types and experimental paradigms. While effects have been discussed in detail in their respective chapters, we will discuss the bigger picture emerging from the patterning of results below, and discuss findings with regards to open questions and future research.

6.1.1 Disgust Sensitivity & Political Values

To the best of our knowledge, this is the first time that Disgust Sensitivity has been investigated in regards to automated language processing, and, as discussed in detail in Section 5.6, it was indeed found to influence language comprehension rapidly. As predicted, Disgust Sensitivity did *not* modulate the comprehension of morpho-syntactic errors or semantic anomalies, but it did affect the processing of socio-cultural clashes (refer back to Section 5.5.3). The direction of the effect was as anticipated as well: Individuals more prone to feeling disgust, and thus more prone to outgroup stigmatization (cf. Aarøe et al. 2017; Faulkner et al. 2004; Inbar et al. 2009; Murray and Schaller 2016; Schaller and Neuberg 2012; Smith et al. 2011), were found to experience greater cognitive load upon encountering a statement clashing with established gender stereotypes.[2] This is in line with the Behavioural Immune System theory, which proposes that individuals more prone to feelings of disgust are also

[2] Note that, at this stage, we cannot make any detailed claims as to what precisely a larger increase in pupil size indicates; as noted in the Introduction, significant changes may be traced back to significantly larger processing difficulty, or a significantly different affective response. More research into the nature of these different aspects of cognition, and whether there potentially may be an underlying variable, is needed.

more likely to have negative feelings towards people identified as "the other" regarding, for example, their customs. Thus, the pupillometric results with regards to Disgust Sensitivity can be linked back to the Behavioural Immune System attempting to allocate attention to the perceived threat (potential pathogen contamination by an out-group, in this case), to try and mitigate its impact on the organism via avoidance.

Disgust Sensitivity has been associated with a conservative leaning in the literature (cf. Aarøe et al. 2017; Murray and Schaller 2016; also refer back to Section 1.1.2.4); however, only a weak correlation was found in the fourth experiment in this dissertation ($r = 0.23$; see Table 1.2), and no significant effect of a conservative leaning on pupil size in response to socio-cultural clashes was found in the fourth experiment. In fact, the opposite effect was found in Experiment III, in which it was *progressive* listeners that experienced a larger change in pupil size when encountering a socio-cultural clash (cf. Fig. 4.3d). This suggests firstly that a listener's Disgust Sensitivity and political values are not directly correlated in all cases (even though some prior literature has found such a correlation; see, for example, Inbar et al. 2009, 2011; Smith et al. 2011), and that secondly, as already mentioned previously, participant sampling may be important for pupillometric results; we will turn to a broader discussion of participant sampling in detail further below. Comparing the time-course of these two effects, it appears that the effect of Disgust Sensitivity sets in quite a bit earlier[3] than that of political outlook (250ms vs. 500ms; compare Figs. 4.3d and 5.4e). This "staggered" effect certainly makes sense, considering that an individual's political outlook – even if potentially correlated with their Disgust Sensitivity in some cases – is comprised of far more than *just* Disgust Sensitivity. An individual's upbringing, education, environment, social circles, volunteer activities, and many other variables influence political leaning, such that it can change over time. This could happen, for example, in a changing environment, when the makeup of an individual's circle of friends changes, or even through deliberate un-learning to fit in with a new social group. Disgust Sensitivity is a comparatively more "primal" variable (Neuberg et al. 2011; Smith et al. 2011) that protects an organism from harm and threat, and that is under less conscious control than, for example, an individual's voting behaviour, or how one chooses to present themselves politically to strangers or acquaintances.

[3] At the same time, it should be noted that the pupillometry paradigm is not as well-suited to the detailed analysis of the time course of an effect as, for example, EEG.

The listener's political values influenced comprehension across all three clash types and experimental paradigms. It was by far the most pervasive extra-linguistic variable in the ratings experiment, influencing ratings across all clash types (refer back to Table 6.1, and Figs. 2.1d, 2.3c, and 2.4d in Section 2.5), and it was correlated significantly with changes in pupil size in response to semantic anomalies in both pupillometry studies (cf. Figs. 4.2f and 5.3c), and in response to socio-cultural clashes in Experiment III (cf. Fig. 4.3d). In most cases, that is, for effects in Experiments I and III, it was progressive individuals that rated clashing items lower, and that experienced a larger pupil size when encountering a clash. As was already discussed in the respective results and discussion sections (refer back to Sections 2.5 and 4.5), this result is not immediately accessible, as it would generally be conservative individuals that would be expected to be "thrown off" more, or experience more surprisal, when coming across less well-formed stimuli. However, investigating the results in more detail, it becomes obvious that progressive listeners simply seem to have a wider scale on which they rate items (see, for example, Fig. 2.1d). Of course, the pupillometric results cannot be explained along those same lines, as they do not involve conscious item ratings. It is possible that progressive individuals experience greater cognitive load when encountering a semantic anomaly or socio-cultural clash due to greater intentional attentional engagement with the stimulus than their conservative peers (Winn et al. 2018). However, in the final experiment, it was *conservative* listeners who experienced greater cognitive effort when encountering a semantic anomaly (cf. Fig. 5.3c), rather than progressive listeners (cf. Fig. 4.2f).

It should be noted in this context that the two political questionnaires, while the first (used in Experiments I through III) includes most of the questions in the second (the Wilson-Patterson test, used in Experiment IV), differed slightly in their structure: The former was made up of two parts, with the first part asking participants to rate their agreement with rather broad statements of a Wilson-Patterson type. However, the second part garnered responses to more detailed questions, to which the Wilson-Patterson test used in Experiment IV has no equivalent. These statements assess world view more broadly, beyond just political values; it may thus be the case that "conservative" and "progressive" mean different things on the two political scales. However, as there was no participant sample which was administered both tests, we cannot make any statements as to whether one test "pushed" participants more towards the ends of the continuum compared to the other, for example. The usage of two different political values questionnaires is a shortcoming of this dissertation, but one that could easily be amended in future research; we will return to this in Section 6.3.2 below.

An interesting observation was the stark difference in timing of the individual difference effects between the two meaning-related clash types: Whereas for semantic anomalies, political values were found to significantly influence pupil sizes around 550ms and 700ms (refer back to Figs. 4.2f and 5.3c), the effects of political values and Disgust Sensitivity on the processing of socio-cultural clashes surfaced much earlier, namely around 250ms (cf. Figs. 4.3d and 5.4e). This timing, in which socio-cultural errors trigger significant changes in pupil size earlier than semantic anomalies, supports the notion that world knowledge, such as aspects of the speaker's identity, and the listener's personal world views, are not considered after the internal semantics and syntax of an utterance have been comprehended. Instead, it seems that, when the speaker's identity has an immediate bearing on the pragmatics of an utterance, individual difference variables influence comprehension even earlier than when comprehending a clash that does not necessarily require access to the speaker's inferred identity.

There is no immediate, clear explanation for why the patterns of results between Experiments III and IV do not converge, especially considering that the paradigm and methodology were the same between the two experiments. Even though the sampling strategy was different in Experiment IV, where about 40% of participants were recruited externally, we cannot conclude that the participant samples in the two experiments differed significantly from each other regarding their political views, as two different political values tests were used in the two studies.[4] However, *within* Experiment IV, externally recruited participants showed a tendency towards a wider distribution of political leanings than the undergraduate linguistics sample (refer back to Fig. 5.1). At the same time, the participant sample in the final study, where recruitment was expanded to include non-academic participants, may differ significantly from the samples in the other experiments *beyond* what can be captured in a single score assessing political outlook. For example, frequent exposure to political discussions or arguments on social media, frequent participation therein, or active involvement

[4] As mentioned previously in the context of the effect of political values, the inability to compare the political questionnaires directly between the two pupillometry studies is a shortcoming of this dissertation, and one that future research should try to address; we will return to this, and potential ways to amend this situation, in Section 6.3.2 below.

with a party or volunteer organization may contribute to a significant difference between individuals that are considered "the same" as per their political values score. Granted, this is (potentially) true for any two participant samples, and is not limited to this dissertation; future research may want to control for such, more "fleeting," and less standardized, variables – within reason, of course, as there is theoretically no end to variables that could be considered influential in experimental research. We will return to further ideas regarding participant sampling and assessment further below.

Another possible reason for the diverging effects of a listener's political values could be that participants were necessarily invited to the lab on different days, at different times, and greeted by different researchers; this may have introduced unintentional biosocial and psychosocial experimenter effects, such as effects of the researcher's appearance or mood on participant performance (Chapman et al. 2018; Rosenthal 1976; Sheldrake 1998). While this is true for all experimental studies, effects may surface especially when investigating responses that involve identity, or responses toward an out-group. Additionally, as several months had elapsed between the participant sessions of Experiments III and IV, the political landscape and the participants' personal experiences with this political landscape may have shifted sufficiently to result in differences in experimental results. More generally, further research is needed with regards to the effects of an individual's political values on language comprehension while controlling for more related variables; some ideas are noted further below. A more speculative explanation for these opposing results could be that, in a politically ever-polarizing world, where online trolls commonly swarm posts that they see as attacking their own views, listeners may have become more sensitive even to statements that they agree with, and hence experience a strong internal reaction even to those items, in a move to "defend" their position from a predicted onslaught from the other side. This is speculation, of course, even if based on personal experience online; however, it would be interesting for future research to investigate physiological responses while individuals read or respond to (polarizing) tweets from both ends of the spectrum, while controlling for the participant's extent of interactions with online social media, specifically in relation to current politics.

6.1.2 Personality Traits

Beyond Disgust Sensitivity and political values, it was the listener's Extraversion and Openness scores that emerged as significant extra-linguistic predictors across experiments and clash types much more frequently than, for example, Conscientiousness (refer back to, for example, Section 2.5.1 and 4.5.1, or 2.5.2 and 4.5.2). We will now discuss these effects in turn.

Comparing the nature of these three Big Five traits, it seems that Extraversion and Openness share a common denominator in different facets of outgoingness – how much joy an individual derives from socializing with others, for example, or how open they are to trying something new. This suggests that, as discussed previously, exposure to non-canonical linguistic stimuli might play an important role here: A more extraverted individual may simply have been exposed more to "strange" statements than an introvert, simply via the opportunities to socialize that they routinely engage in. In turn, this would then suggest that, as per a vaguely Bayesian or experiential approach,[5] it is not just the personality trait *per se* that influences language comprehension, but that it may be past experiences that the individual engaged with differently as per their personality that now affect language comprehension. That is, it is not necessarily the trait as such that modulates how someone understands a statement; but rather that this personality trait modulates the way that an individual engages with and navigates the world, thus leading to different experiences with linguistic stimuli, which then in turn modulate linguistic comprehension.

While there is a lot of overlap in effects between the ratings and pupillometry experiments for morpho-syntactic errors and semantic anomalies (recall that, for example, Openness was found to be a significant predictor in modelling the comprehension of morpho-syntactic errors in both the ratings and the second pupillometry experiment, and that Extraversion was a significant predictor for semantic anomalies in the ratings and first pupillometry experiment; see also Table 6.1), there are some discrepancies in the directionality of the effects in the two paradigms that cannot be explained conclusively at this time. For example, more introverted listeners rated semantically anomalous items worse, and experienced greater cognitive load upon encountering such an item, than their more extraverted peers (refer back to Figs. 2.3b and 4.2d); this was expected, and both effects point in the same direction: Introverted listen-

[5] We will turn to a (brief) discussion of theoretical accounts in this context further below, in Section 6.2.

ers seem to take more issue with a semantic anomaly, both consciously and subconsciously. The effect found for socio-cultural clashes in Experiment IV points in the same direction, with introverted listeners experiencing greater cognitive load when coming across the clash (cf. Fig. 5.4d). As alluded to in previous sections, this may conceivably stem from the fact that introverts simply have not been exposed as much to "odd" utterances due to fewer (or less diverse) social interactions, and thus anticipate canonical utterances more than their extraverted peers. However, the image that emerges for morpho-syntactic errors – the only type of clash that does not draw on meaning or world knowledge to be a clash – is different: Here, it was more *extraverted* listeners who rated utterances containing an error worse than did their introverted peers, with *introverted* listeners experiencing greater cognitive load when encountering a morpho-syntactic error. There is a discrepancy here, in which extraverted listeners seem to take more conscious issue with a morpho-syntactic error, whereas introverted listeners seem to experience greater cognitive load as they encounter the error – but do not let that influence their conscious item ratings (refer back to Table 6.1). These results should be interpreted in conjunction with findings from Boland and Queen (2016), specifically Fig. 3 (bottom) on p.10, where introverts rated the author of an email as a worse housemate when the email contained typos; for extraverted individuals, the difference in ratings was not nearly as pronounced. This shows an interesting discrepancy, which may have to do with the difference in error type (typo in Boland and Queen 2016 vs. morpho-syntactic agreement error in this dissertation), but which also may have to do with what exactly the experiment assessed: Whereas in the ratings study in this dissertation, participants were asked to rate the utterance itself for acceptability, in Boland and Queen (2016) they were asked to rate their potential future housemate, that is, *a person and the quality of their character*. Thus, in addition to the difference between sub-conscious, immediate processing (such as assessed via the pupillometry paradigm) and conscious processing after the fact (as in the ratings paradigm), an additional "third" level may be that of adjusting one's behaviour, or judgment, in a social context. While an individual may have a strong gut reaction to an error or anomaly, and may even consciously rate the error or anomaly as annoying or not acceptable, they may be able to "override" these reactions when extrapolating from the error or anomaly to, for example, another person's character. The extent to which this extrapolation is or is not happening may then be affected by the listener's personality. More research is needed here, to assess how precisely the three "levels" of comprehension differ from one another, and how certain personality traits specifically modulate whether an initial gut reaction propagates to the more conscious, active levels of assessment. For example, future research could include another type of intra-linguistic error in addition to morpho-syntactic agreement errors, and assess subconscious, immediate processing (such as via pupillometry or EEG experiments), conscious item ratings, *and* an assessment of the character of the person that produced the error. In either case, results do suggest that even the "first contact" with the content of an utterance is modulated by the listener's personality, in line with a one-step model of language comprehension; the question (to tackle for future research) alluded to here is whether this modulation changes over the course of time, when the more conscious aspects of interpretation kick in.

Looking now at the pattern of effects related to the listener's Openness, a clear divide is visible between morpho-syntactic errors and semantic anomalies on the one hand, and socio-cultural clashes on the other (cf. Table 6.1): Whereas for morpho-syntactic errors and semantic anomalies, it was *more* open individuals that rated clashing items worse and that experienced a greater cognitive load when coming across the clash, it was *less* open listeners who rated utterances with socio-cultural clashes worse than their more open peers, and who also experienced more cognitive load when they encountered a clash of this type. This very much again highlights the interesting in-between status of semantic anomalies that was already discussed in Section 4.6: A statement like "they often read *heads* for pleasure at night" is anomalous, as generally heads cannot be read; however, embedded in the right (fictional) context, where for example text is commonly printed on bald head statues (recall the peanuts example from Nieuwland and Van Berkum 2006), such a statement *could* be non-anomalous. As such, semantic anomalies do rely on an individual's prior experience with the world, and hence are not quite as clear-cut as morpho-syntactic errors; however, socio-cultural clashes go a step "further" as they derive their strangeness *exclusively* from a clash with the speaker's inferred identity. As already discussed, it seems that more open individuals (much like more progressive individuals) simply seem to have a wider scale on which to consciously rate utterances; their interactions with the world may have given them a broader "scope" which they use to categorize stimuli. However, this explanation is not sufficient for the pupillometric results, as those do not involve conscious rating or decision-making. It is possible that more open listeners, akin to more empathetic individuals, or individuals in a good mood, may engage in more linguistic anticipation based on the information available

(Havas et al. 2007; Van Berkum et al. 2013; Van den Brink et al. 2010; Zadra and Clore 2011; also refer back to Section 1.2.2), which then surfaces as greater processing load if a morpho-syntactic error or a semantic anomaly causes those anticipations to not be fulfilled. In the case of socio-cultural clashes, it is then further possible that open listeners, even if they are subconsciously anticipating no clash, are simply more comfortable with statements that do not agree with established gender stereotypes. Thus, greater Openness may render an individual more comfortable with clashes related to the inferred identity of the speaker, but not necessarily with violations that do not draw on the identity of an individual.

While the two pupillometry experiments in this dissertation did not investigate changes in pupillary responses over the course of the experiments, Van den Brink et al. (2010) noted a difference between semantic anomalies and socio-cultural clashes: The N400 that was observed for both semantic anomalies and socio-cultural clashes in the first block of their experiment remained in the second block *only* for semantic anomalies. The authors reason that socio-cultural priors may be easier to update than semantic ones in a Bayesian model. While this cannot be extrapolated with certainty, it is possible that this difference in updating priors is responsible for the difference in the effect of Openness on the processing of semantic anomalies vs. socio-cultural clashes in this dissertation: More open individuals may have been more successful at updating socio-cultural priors, so that less open individuals experience greater cognitive load when encountering a socio-cultural clash, when it was more open individuals that experienced greater cognitive load when encountering a semantic anomaly (refer back to Table 6.1). However, note that this dissertation does not at its core pose the question of whether Bayesian reasoning is involved in language comprehension, so this discussion is speculative at this point (but may be of interest for future research that is more directly aimed at this particular question).

6.2 Speaker-Related Variables

In addition to listener-internal variables modulating language comprehension, information inferred about the speaker – specifically, their gender – was found to affect language comprehension as well. This happened in one of two ways: The speaker's gender was either found to be a significant predictor in the model with the best fit (see, for example, Table 3.1 and Fig. 3.1c, where response times were generally slower when the speaker was male); or it was explicitly a part of the clash, as in the socio-cultural clash condition, where clashes relied on established gender stereotypes. Importantly, in this latter scenario, the gender inferred about the speaker, based on their voice, *interacted* with the listener's internal state (such as their personality or political views) to affect language comprehension. For example, as was shown in Experiment IV, less open individuals were found to experience a higher cognitive load upon encountering a statement that clashed with the gender inferred from the voice of the speaker (cf. Fig. 5.4c); the same was found for individuals with higher Disgust Sensitivity (cf. Fig. 5.4e). The results thus support the notion that information about the speaker is inferred from their voice and used as a clue (Belin et al. 2011; Ko et al. 2006), and that the way this information is used is modulated by certain listener-internal variables. So, as expected from prior research discussed in the Introduction (specifically refer back to Sections 1.1.1.2 and 1.1.1.3), it thus indeed seems that the extent to which a listener relies on stereotypes based on voice cues is affected by their own internal state. This is in line with results from Van Berkum et al. (2008), where segments not corresponding to the gender inferred about a speaker triggered significantly different ERP signatures; and with Quadflieg and Macrae (2011), where more prejudicial individuals seemed to activate stereotypical information more than their less prejudicial peers. With many of the observed extra-linguistic effects surfacing around 250 or 300ms (cf. Figs. 4.2d, 5.4c, or 5.4d), results are in line with significant ERP effects reported in the literature (cf. for example the N400 effects found in Nieuwland and Van Berkum 2006; Van Berkum et al. 2005), especially considering that pupil size is a more slow-moving measure than EEG measurements. Socio-cultural clashes specifically were associated with a significantly larger increase in pupil size around 300 to 400ms after clash onset, that is, the same time when semantic anomalies triggered a significantly larger pupil size as well (refer back to, for example, Figs. 4.2b and 4.3b). Stereotypes thus seem to influence language comprehension right as it happens; information about the speaker seems to be integrated right away, and the extent to which this happens seems to be modulated by listener-internal variables.

While not at the heart of this dissertation, linguistic research relating to stereotypes could have implications for the organization and representation of stereotypes themselves. A detailed discussion of current models of stereotype representation is beyond the scope of this work; however, three different model types will be discussed briefly.[6] Early prototype research favoured abstract, **prototype-/schema-based accounts** (Cantor and Mischel 1979; Coats and Smith 2007; Johnston and Hewstone 1992; Park et al. 1991). In a model of this type, an abstract average of the typical features of a group is stored, and an individual that is encountered is then subsconsciously compared to this average (Hilton and Hippel 1996). Prototype-based models are considered rather stable over time, and tend to not be influenced by different contexts or individual differences (Coats and Smith 2007). In an **exemplar-based model**, on the other hand, a number of distinct exemplars are stored, and represent a group. Individuals are not compared to an abstract average, but rather to the unique features of specific group members, even if each of those group member is a "far cry" from the group's stereotype (Linville et al. 1989; Smith and Zárate 1992). This type of model is considered less stable, more fluid, and more susceptible to social context and individual differences, such as past experiences or motivation at the time (Potter 2002). For example, research suggests that attention may be given to one dimension of an individual (like gender) over another (like race), depending on the context (Smith and Zárate 1992). Coats and Smith (2007) suggest that the inherent flexibility in stereotype retrieval, for example between members of an in-group versus an out-group, can only be accomplished through information stored in exemplars. A counter-stereotypic exemplar may be able to change the stereotype quite rapidly (Hilton and Hippel 1996); the view one has of a particular group may change from one situation to the other, depending on which specific exemplars are recalled (Coats and Smith 2007). Purely prototype-based models are not generally thought to allow for such flexibility. As an "in-between" approach, **mixed models** borrow aspects from both prototype- and exemplar-based accounts. They propose that, in one particular context or situation, one may rely more on abstract information, whereas in another context, information stored in exemplars may be used more significantly (Coats and Smith 2007). Research suggests, for example, that representations of in-groups rely more on exemplars, whereas representations of out-groups rely more on a prototypical representation (Coats and Smith 2007; Park et al. 1991). Some mixed models, such as the varying abstraction model, consider prototype

[6] Note that, while the following discussion of abstraction in cognitive representation focuses on general cognition, the same distinctions are present in different competing linguistic theories; experiential and exemplar-based, rich memory accounts in a linguistic context will be discussed further below.

and exemplar theory the ends of a continuum, and allow for partial abstraction (Vanpaemel and Storms 2008, 2010). Mixed models thus also allow for flexibility in how groups are represented, and how these representations are accessed under different conditions.

The results of the four experiments reported above cannot clearly distinguish between these three approaches to stereotype representation. However, given the demonstrated influence of individual differences (such as political views, Disgust Sensitivity, and personality traits) on the processing of socio-cultural clashes, results lean more towards an approach that can easily accommodate the influence of context and individual differences, such as an exemplar-based or mixed model. To make a more definitive claim, more research with a modified paradigm would be needed that specifically assesses the variability of stereotypes in language comprehension. For example, participants could be presented with images of different popular members of a gender category before a block of trials, or before the start of the experiment. These category members should be selected based on differing perceived distances from the prototype – say, a very feminine woman, as opposed to a woman that has stereotypical masculine traits, skills, or appearances. By observing the influence of those different group members on subsequent responses (ratings, response times, or changes in pupil size), a more detailed claim regarding how stereotypes seem to be represented, and how these representations affect language processing, could be made.

Within the context of both extra-linguistic information, and abstraction in memory, the literature on grounded cognition and experientialism must be acknowledged. The body of research is, at this point, substantial, and makes distinctions in a nuanced way between *embodiment*, *grounded cognition*, and *experientialist accounts* (see, for example, the discussion of terminology in Barsalou 2008). We will not discuss these fine distinctions in detail here, as the four experiments in this dissertation cannot help inform the discussion around whether language comprehension (or how much of it) is grounded, embodied, or rooted in experiential representation. At the same time, experientialist accounts do overlap in parts with extra-linguistic information; for this reason, some of the existing literature will coarsely be reviewed here, "lumped together" under *experientialist accounts* for the purposes of this dissertation.

Experiential information is, by definition, extra-linguistic. It is information gathered by interaction with the real world – such as properties and affordances of objects, and physical and mental states associated with a word or experience (Andrews et al. 2009; Kaup et al.

2007). As Barsalou (2008) notes, experiential accounts differ from standard cognitive theories in that they do *not* assume that the brain stores, and operates with, amodal symbols, but rather that perception, action, and introspection form the basis for knowledge. Within an experientialist framework, language comprehension is thus considered a part of ordinary, general cognition (Bybee 2010; Gibbs and Perlman 2010), where words are stored along with traces of perceptions and actions that are associated with them (Barsalou 2008; Harris et al. 2003; Zwaan et al. 2004). Abstract concepts, such as time, knowledge, or love, are assumed to be understood via more tangible human experiences (Boroditsky et al. 2001; Gibbs 1994; Lakoff and Johnson 1980, 1999; Matlock et al. 2005). Based on experimental research investigating language comprehension, there is evidence within the experientialist literature that listeners simulate the state or action that is being described (Kaup et al. 2007); or, more generally, that perceptual representations are activated during language comprehension (Zwaan et al. 2004).[7] This ties in with abstract concepts being understood in terms of more discrete concepts, as just mentioned previously: For example, the perception of time (a rather abstract concept) seems directly modulated by a person's perception of space (a more tangible, "relatable" domain), to the point where the perception of time is modulated by how far back in a lineup someone is waiting, or whether they envision themselves "moving through time," or time moving around them (Boroditsky et al. 2001; Matlock et al. 2005). Results from Glenberg and Kaschak (2002) and Zwaan et al. (2004) suggest that simulations of movement are involved in the comprehension of sentences describing motion. In this context, also recall the findings of Havas et al. (2010, 2007), discussed in Section 1.1.2.2 in the Introduction, where sentences were judged faster when an induced facial configuration – a changed bodily state – matched the valence of the sentence. This is, of course, just a very small sample of experientialist research; the reader is directed to Barsalou (2008) for an excellent review of the existing literature in the field, and Kaup et al. (2007) in particular for relevant studies investigating language comprehension from an experiential viewpoint.

As discussed previously, the experiments in this dissertation do not, by themselves, support or contradict experientialist accounts. At the same time, any research on the influence of extra-linguistic information on language comprehension should acknowledge that, at the very least, there is common ground with experientialist accounts in the rejection of a two-step

[7] In this context, also note the (very) tangentially related work on the importance of memory for adaption and for imagining the future (Klein et al. 2002; Schacter et al. 2012).

approach to language comprehension: As Gibbs and Perlman (2010, p.3) put it, “there is no evidence that people automatically create literal, semantic, purely propositional representations for sentences (i.e. a ’sentence meaning’) before elaborating on these representations to infer speakers’ and writers’ broader communicative messages (i.e. a ’speaker meaning’).” Theoretically, it may well be possible for a statement or a phrase to be stored as a trace that also references the feelings and “gut reactions” of a listener at the time. For example, if a phrase like “I often wear a dress to work” is only ever uttered by stereotypically female speakers, this factoid may be included in the memory trace. Then, if the phrase is produced by a stereotypically male speaker, the extant association with stereotypically female features stored alongside the memory trace may become activated and cause a clash. This is, of course, speculation, and more research is needed here; future research may want to investigate the influence of extra-linguistic information on language comprehension specifically in relation to experientialist accounts – it is well possible that an individual’s personality, political values, or Disgust Sensitivity may modify how traces are (or are not) filtered by selective attention (Zwaan et al. 2004), and how they are subsequently stored as knowledge; or, alternatively, how perceptual traces are retrieved when a triggering stimulus is encountered. Much more research is needed here, for which extra-linguistic variables, such as those identified in this dissertation, may be an interesting and appropriate testing ground.

As a final note on experientialist accounts in the context of extra-linguistic information, its relationship to Bayesian processing should be mentioned briefly. While experiential accounts seem to be diametrically opposed to distributional or statistical accounts of meaning at first glance – one derives meaning from interaction with the world, whereas the other remains on a purely intra-linguistic level and derives meaning from how a word patterns with other words – Andrews et al. (2009) found that, when experiential and distributional information were combined, the modelling of semantic learning was greatly improved. This is very broadly in line with a Bayesian approach, in which listeners use *all available information*, intra-and extra-linguistic, to estimate the likelihood of an upcoming segment (Traxler 2014; see also Barsalou 2008; Clark 2013). Although beyond the scope of this dissertation, considering a statistical approach in conjunction with an experiential viewpoint, rather than as an opposing theory, may be a promising path for future research to take.

6.3 Miscellaneous Findings & Future Research

As is the case with every study, some shortcomings and new questions were identified in the previous chapters, for example regarding participant sampling and assessment. Various ways in which future research could improve on these shortcomings, and/or drill down to a specific aspect of individual differences in language comprehension, will be presented below.

6.3.1 Methods & Paradigms

Comparing effects across the three experimental paradigms, it must be noted that the self-paced listening study overlaps the least with the effects found in either the ratings or pupillometry studies. This can likely be traced back to a slightly different model structure that was used to model response times: As the pupillometry models were already rather large and complex, which is not all too desirable for a GAMM, a decision was made to not investigate complex interactions that included gender variables at this time. However, this type of interaction was investigated in the modelling of SPL responses, so that results cannot immediately be compared in regards to the presence or absence of explicit gender interactions. Even beyond complex interactions with a gender variable, however, there was little overlap in significant listener-internal variables between the SPL study and the other three experiments (refer back to Table 6.1): In fact, it was only Conscientiousness that was found to interact significantly with item condition, and only in the modelling of response times to morpho-syntactic errors – a variable that is markedly absent from all other studies in this dissertation. At the same time, this lone Conscientiousness effect, in which item rating had no effect on the response times for more conscientious listeners, but only for less conscientious listeners (cf. Fig. 3.1a), is an interesting finding which suggests that response times in a self-paced listening task may be strongly influenced by the participant's conscientiousness level. For example, an effect might only surface for those listeners that are less conscientious, whereas higher Conscientiousness scores may "mask" any effects, as the participant makes an effort to pay close attention to the experiment, and respond quickly in all cases, regardless of item condition. The distinction between these two groups would be invisible to the researcher unless participants had been assessed for their Conscientiousness. As such, the SPL paradigm – and, by extension, other paradigms using conscious measurements, although more research is needed there – should be used with caution if participants are not administered a personality test that assesses Conscientiousness, as crucial effects may be masked. However, if the goal is to assess different facets of language comprehension in an array of different paradigms, an SPL experiment can be highly useful precisely to identify personality traits that affect conscious measurements, as opposed to changes in pupil size or ERP signatures.

Future research should also explore additional experimental paradigms, specifically EEG, and potentially even a co-registration paradigm between ERP and pupillometry. This would make findings directly comparable to the body of research already in existence regarding the effect of empathy on language comprehension (such as Van den Brink et al. 2010), and it could reveal important information regarding when exactly an effect happens. Co-registering pupil size and EEG for written stimuli is highly problematic, as the reader will need to fixate the text; however, as all four experiments reported here used auditory stimuli, and as data can be gathered without the participant having to complete a task, co-registration is generally possible.

6.3.2 Participant Sampling and Assessment

As was noted throughout Chapter 5 and in earlier parts of this General Discussion, the sampling strategy in the final experiment in this dissertation differed slightly from the usual undergraduate student sample that was used for the first three experiments. A comparison between the two different groups in Experiment IV showed that there were significant differences only for the Neuroticism and Openness trait, where the externally recruited participant group was more open and less neurotic on average than students recruited from the undergraduate pool. Note that it is at this point unclear whether this difference in samples would extend further, that is, to a difference in the underlying populations; as discussed in Section 5.1, it is possible that the external sample is self-selecting in nature, such that the sample may end up being less neurotic and more open than the population it is recruited from. Comparing significant listener-internal variables between the two pupillometry experiments (refer back to the two rightmost columns in Table 6.1), it is immediately evident that Experiment IV did not replicate all of the effects found in Experiment III. In fact, more effects *differ* between the two experiments than are shared between them – the only replicated effect is that of Extraversion on the processing of morpho-syntactic errors. At the same time, many of the effects that changed between the two studies (Openness and

Neuroticism for semantic anomalies, or Openness for socio-cultural clashes, for example), relate back precisely to the two variables whose distribution differed between the two recruitment groups. While the overall trait distributions did not differ significantly between Experiment IV and all other experiments in this dissertation (refer back to Table 1.1), the correlations observed between the different traits differed within each participant sample (cf. Table 1.2). This suggests that, even if two or more participant samples are highly similar in the overall distributions of personality traits, the individual combinations of traits, and thus the resulting correlations as well, can be rather different. Experimental results may then differ due to the differing underlying participant samples, even though the two samples look the same with regards to the overall trait distributions. This highlights a range of concerns related to participant sampling that future research may want to address. Namely, the results discussed highlight the importance of:

1. Sampling from the general population while at the same time being wary of self-selecting circumstances;
2. Assessing personality trait correlations in the sample *beyond* a simple distribution; and
3. Controlling for as many individual differences as possible, with the goal to compare results between samples that differ in a small number of aspects.

While items 1 and 2 have already been discussed above, we will now turn to a more detailed description of item 3. Ideally, future experiments should consider and assess as many individual difference variables as possible, so that participant samples can be compared along just a few differing "dimensions." For example, take two participant groups that differ significantly in their age distribution and in their Extraversion and Openness scores. Even though it is possible to control for these differences via predictors in a statistical model, this has the potential to overcomplicate the model, which is a concern as it limits the amount of deductions that can be made; and adding too many predictors may make the model not converge at all. Instead, participant samples could be recruited in such a fashion that the samples differ significantly in *either* their age distribution, *or* the distribution of a personality trait. Thus, by attempting to keep other variables constant, stronger claims could be made regarding the influence of one particular variable. Note that, while this approach sounds highly desirable in theory, it may be very difficult, or even impossible, to achieve realistically when recruiting participants for an experiment due to constraints (whether monetary or logistical) in the real world. In addition, as has hopefully become clear over the course of this dissertation, it seems that more individual difference variables that affect language comprehension are being discovered. Thus, it is at this point not possible to tell which other variables may be out there that influence language comprehension that we are not currently aware of, and that should be controlled for. A reasonable approach for future research may thus be to assess the participants' personality, and to stay aware of other variables that have been found to affect language comprehension or related aspects of general cognition in recent research, so that they may be considered as predictors (see also further below for more specific ideas on additional variables to consider). This way, a "body" of research, and a catalogue of variables, could be compiled; new variables can then incrementally be compared and added to this body of research, and further correlations between variables can be established.

While the Big Five test has performed well as a personality assessment in the four experiments, it may be worthwhile to explore the *HEXACO* assessment, which includes an additional trait – the *H factor*, or *Honesty/Humility* – that has been found to significantly influence various aspects of general cognition in recent research (Ashton and Lee 2007; Lee and Ashton 2004; Parks-Leduc et al. 2015). Importantly, Tybur and Vries (2013) have found links between Disgust Sensitivity and the added Honesty/Humility factor, a link that cannot be captured by the Big Five assessment. However, even when adjusting the personality assessment, an important caveat remains: As alluded to in several sections of this dissertation, the four experiments and their results cannot help distinguish between effects that relate directly to a listener-internal trait, versus *effects related to differences in prior exposure to linguistic stimuli* that were caused by the trait in question; or, as Traxler (2014) puts it, "Individual differences may also be found in the way that comprehenders acquire the knowledge that drives estimates of prior probability" (p. 610). While this difference is notoriously difficult to assess empirically, and strongly relates back to the underlying theoretical debates of whether representation is exemplar-based/experiential or not[8] (and thus should tie in with new research and developments in those specific areas), some simple additions to the language background questionnaire (such as "How many hours a week on average do you spend at parties, gatherings, or otherwise socialize with friends or colleagues?") may be a good first step.

[8] See Section 6.3.5 below for a discussion.

More generally, attempts should be made in future research to assess more types of individual differences for each participant – beyond personality traits. Crucially, if socio-cultural clashes based on established gender stereotypes are used as stimuli, participants should be assessed regarding their sexual orientation. Especially when investigating the influence of Disgust Sensitivity, it would be important to be aware of what sexual orientation the participant identifies with, as this may change the participant's relative perceptions of in- and out-group status, and thus the responses triggered by certain stimuli. For example, a bisexual individual might be less at odds with stimuli referring to homosexual acts, even if their general Disgust Sensitivity may be significantly higher than that of the average population. It may also be interesting to tease apart which specific aspect of disgust these influences stem from in particular; thus, future research may want to assess Disgust Sensitivity not just along one general scale, but rather its three subscales of pathogen disgust, sexual disgust, and moral disgust. These three dimensions of Disgust Sensitivity may interact in meaningful ways with stereotypes regarding features other than (binary) gender, as discussed above. Additionally, said Disgust Sensitivity was assessed in only a static fashion for each participant in the final experiment in this dissertation. Since it has been shown that an individual's disgust sensitivity is not static over time, but fluctuates given information in one's surroundings and can thus be manipulated (Helzer and Pizarro 2011; Schaller and Neuberg 2012; Schaller and Park 2011), an experimental manipulation of perceived pathogen threat, such as through reminders of a flu epidemic, or to wash one's hands frequently, may give interesting insight into the functional flexibility of the BIS *within* individuals.

Further additional aspects of the participants' experience in the world, that may conceivably influence their response to certain clashes, were not captured in the statistical models reported above. It is conceivable that, for example, whether an individual likes to see shows at a theater, or likes to watch movies that challenge their world view, listen to world music, or explore different literature genres, may very well affect how they navigate the world and comprehend language, without this being captured in a standardized personality test or a test of political values. Further variables such as the makeup of friend groups, and what causes an individual volunteers their time for, may also distinguish one person from another that has the same Big Five trait combination. In addition, this may mean controlling for (either by assigning participants to different groups while running the experiment, or by controlling for it statistically in the models after the fact) which researcher greeted the participants in the morning, and the researcher's gender and personality, so as to avoid unintended biosocial and psychosocial experimenter effects (Chapman et al. 2018; Rosenthal 1976; Sheldrake 1998). It may also be worthwhile to assess the participant's mood as they come in to the lab, as mood has been shown to affect linguistic processing (Havas et al. 2010, 2007; Van Berkum et al. 2013), but is unaccounted for in a personality measurement. Of course, experiments between which results are to be compared should use the same political test; this is an unfortunate shortcoming, especially between the two pupillometry experiments in this dissertation, but one that is easily amended and would not add "bulk" to the post-tests in any study.

In summary, future research may want to control for more "fleeting" variables – while at the same time not requiring participants to fill in questionnaires for several hours, thereby potentially rendering the experiment highly unnatural.

6.3.3 Stimuli & Speaker Gender

To address the shortcoming in which there was only one speaker per (binary) gender in all four experiments, more speakers per gender should be recruited in future research. This way, effects caused by idiosyncrasies pertaining to each speaker could be avoided. Furthermore, the four experiments in this dissertation treated speaker gender as a binary variable – it would be worthwhile to recruit several different speakers, and have their voices rated for masculinity/femininity in a separate study, thus placing their voices on a gender spectrum or continuum rather than in binary categories. This variable could then be used as a numeric predictor, thus moving beyond a binary view of the gender concept, especially if some speakers are at opposite ends of the spectrum (i.e. stereotypically "male" or "female" speakers), and others are more centrally located (i.e. more ambiguous, or non-binary, in regards to their voice). Broadening the speaker gender feature beyond just one stereotypically male/female speaker each could also form the basis for a more detailed investigation of gender stereotypes – while gender is now seen as a more fluid concept than it was, for example, 20 years ago, old male/female stereotypes are very much still alive. At the same time, new stereotypes have evolved in regards to non-binary, queer, and transgender individuals. A separate ratings experiment that places speakers on a continuum would make an investigation of these wider facets of the gender variable possible, adding more data to an investigation of how stereotypes influence language comprehension, and how they interact with listener-internal variables. Of course, the socio-cultural clash trait could always be expanded to include dimensions other than gender; Van Berkum et al. (2008) and Van den Brink et al. (2010) have, for example, used speaker age and social class as additional dimensions. This dissertation limited socio-cultural clashes to the gender dimension to not overload the models; future research may want to investigate other dimensions systematically, while taking into account the improvements related to participant assessment discussed just above.

6.3.4 Statistical Modelling

While GAMM modelling has been found to be well-suited to the analysis of time series data, the models in this dissertation are rather large and try to do many things at once. Now that general effects of listener-internal and speaker-inferred variables have been established, it may be worthwhile to design future experiments such that they assess the influence of only a few factors at a time. For example, by investigating just the influence of the Honesty/Humility factor in the HEXACO assessment, three-way interactions with listener and speaker gender could be assessed without overloading the model. Likewise, experiments could be designed to only assess one clash type, or even just one sub-type of clash (such as the age or social class dimension just mentioned in the previous section) at a time, to reduce the number of predictors in the model, and thus increase their power and drill down into one particular aspect of language comprehension at a time.

6.3.5 Broader Theoretical Questions

It is important to note that the experiments in this dissertation were not designed to distinguish between different theories in the realm of syntactic parsing or ambiguity resolution – neither of the experiments overtly modulated or assessed the noisiness of the signal, ambiguity in interpretation, or the probability of alternate syntactic interpretations. For this reason, we cannot use the results discussed above to distinguish between different syntactic parsing theories that consider extra-linguistic information, such as noisy-channel processing and good enough parsing (see e.g. Ferreira 2003; Levy 2010; Traxler 2014; Trueswell et al. 1994). At the same time, results are broadly in line with constraint-based theories of sentence processing, in which information from a variety of sources (intra-linguistic as well as contextual, pragmatic, real-world information) are considered without delay, as soon as they are available (McRae and Matsuki 2013; see also Boland and Queen 2016; Kamide et al. 2003; Tanenhaus et al. 1995; Van Berkum et al. 2008).

Discussing the results of the four experiments with regards to broad cognitive theories likewise goes beyond the scope of this dissertation, as mentioned in the Introduction. However, results are generally compatible with a Bayesian approach: Introversion, by shaping an individual's experiences (or non-experiences) with linguistic stimuli, thus may also shape priors that are considered in linguistic anticipation. However, note that results are only partially compatible with a bio-energetic account in which a more extraverted or open demeanour, or a more progressive outlook, akin to a good mood, would result in more cognitive energy being allocated to "going out" and "exploring" (see also Dewaele and Furnham 1999; Havas et al. 2007; Van Berkum et al. 2013; Zadra and Clore 2011). This would then be expected to result in more linguistic anticipation, and thus *more* surprisal at an unanticipated segment, such as one that clashes with the listener's perceived identity. Some effects found across the four experiments are supportive of this theory (refer back to Table 6.1), such as the effect of higher Openness and Extraversion scores on morpho-syntactic error item ratings, or on pupil sizes in response to semantic anomalies in Experiment III. However, a number of other effects are not in line with the theory, such as the effects of low Agreeableness and Openness on the processing of semantic anomalies and socio-cultural clashes. At the same time, recall that research directly related to the bio-energetic account investigated mood and empathy; it may simply be the case that Big Five traits, in addition to an individual's political values, are not immediately related to resource allocation in the same way that empathy and mood may be. Research specifically targeting the bio-energetic account in regards to personality traits and political values in language comprehension is needed here, if the goal is to make an accurate claim about the relationship between the theory and these individual difference variables.

Of course, a dissertation in experimental psycholinguistics cannot really conclude without at least attempting to address what the results mean for linguistic representation, and for language as such. What is language if its comprehension is affected by extra-linguistic individual difference variables, such as the listener's personality and Disgust Sensitivity? As has hopefully become clear from the previous sections of this General Discussion, much more research is needed, with various improvements, small and large, to make a definitive claim as to what extra-linguistic effects mean for language comprehension and linguistic representation. At this stage, we must resist the temptation to try and answer an age-old question definitively based on promising, but still rather early and exploratory, results. The paragraphs below are thus at best an attempt to broadly categorize the findings in this dissertation within larger frameworks, and to assess what they tell us about language.

As mentioned just above, at the beginning of this section, results are generally supportive of constraint-based theories of language comprehension, in which a variety of different types of information affect comprehension early on; in one single step, various sources of information are used to interpret an utterance. As such, language seems to be directly reflective of context – including its non-linguistic aspects, such as the real world which language is embedded in, and the listener's experience within it. Results are thus in agreement with a cognitive linguistic approach, where language comprehension is not assumed to create a context-free interpretation first; results agree with the cognitive linguistic views on how meaning is dependent on non-linguistic context from the start (Gibbs and Perlman 2010; Ibbotson 2013), and how language is inherently shaped by the context in which it is used (Haspelmath 2002; Kristiansen 2006; Trask 1999). As personality and Disgust Sensitivity, both decidedly non-linguistic aspects of cognition, have been found to influence language comprehension in this dissertation, results further support a view in which language is at the very least not a (fully) domain-specific, closed-off system. Again in line with a cognitive linguistic view, which considers language to be a part of general cognition, assigning limited (if not zero) importance to innate capabilities (Croft and Cruse 2004), how humans perceive the world seems to be reflected in how they comprehend language: Someone who is more open or more progressive, for example, seems to have a different "lens" through which they view the world, which also seems to be affecting language comprehension.

Although much of the following is speculative at this stage, as no strong claims can be made in either direction based on the results of this dissertation, results may also be broadly supportive of (at least some facets of) usage-based aspects within cognitive linguistics (Bybee 2010; Geeraerts 2013; Ibbotson 2013). Under this account, memory is considered "rich," holding on to (theoretically) all perceptible types of information, intra- and extra-linguistic, using exemplars (Bybee 2010).[9] Note that this strongly relates back to our earlier discussions of exemplar-based accounts in the context of stereotypes, and our discussion of experientialist views, both in Section 6.2; exemplar representation is considered crucial to a usage-based approach (Bybee 2010; Ibbotson 2013). Results from this dissertation are not incompatible with an account that stores more information than just grammatical properties alongside words or phrases. Note that, however, this possibility does not automatically mean that results agree with further tenets of usage based-accounts, such as their interpretation of grammaticalization, or gradients in place of clear-cut categories – we cannot make any claim in those regards based on the four experiments in this dissertation. At the risk of sounding like a broken record: Much more research is needed here. As usage-based approaches generally assume that incoming information can change stored representations on-the-fly, via exemplars and the memory traces associated with them, one avenue of investigating the intersection of usage-based accounts and extra-linguistic information may be to manipulate incoming traces, for example by having stimuli produced by speakers of varying genders in a training phase, and then assessing how these manipulations affect language comprehension in a subsequent testing phase. Disgust Sensitivity may be an interesting candidate, as it can be manipulated via the presence of perceived pathogen threat, as discussed above.

As a final, very general note on the effects found in the four experiments, it is important to remember that none of the relationships in any of the four experiments are to be taken as causal relationships. That is, simply because political values (as an example) emerged as a significant predictor in a model fit for changes in pupil size, and the phrasing that is commonly used describes a variable "influencing" the results, this should not be taken to mean that political values directly *cause* changes pupil size, or, by extension, changes in cognitive load. It is possible that an underlying variable, such as a combination of Disgust Sensitivity and an as of yet unidentified variable, causes the differences in cognitive load. The experiments discussed above cannot determine this; further research at the intersection of psycholinguistics, brain studies, and biology are needed to tease these effects apart.

[9] Compare, in this context, the notion in Van den Brink et al. (2010) that analyzing the effects of individual differences may help us learn more about the underlying processes than if we were to simply discard them as "noise."

6.4 Conclusion

In summary, based on the results of the multi-methodological array of experiments in this dissertation, and addressing the research questions posed at the outset, we can conclude that:

- Several variables related to the internal state of the listener, and to information inferred about the speaker, influence spoken language comprehension.
- Not all Big Five traits influenced comprehension to the same extent or in the same manner, and the three types of clashes – morpho-syntactic errors, semantic anomalies, and socio-cultural clashes – were affected by different extra-linguistic variables.
- This suggests that a multi-methodological approach may provide a more comprehensive account of the phenomenon in question, as compared to a single-method study (see also Arppe and Järvikivi 2007a,b).
- Crucially, Disgust Sensitivity was shown to modulate the comprehension of socio-cultural clashes.
- Language comprehension is affected by these extra-linguistic variables right as it happens. Results are thus in line with one-step models of language comprehension.
- The findings broadly support cognitive linguistic, constraint-based approaches, with tentative support for some aspects of experiential and usage-based accounts.
- Much more research regarding the influence of extra-linguistic information on language comprehension is needed to make definitive claims regarding the nature of linguistic representation, and the interface of general cognition and language.

References

Aarøe, Lene, Michael Bang Petersen, and Kevin Arceneaux (2017). "The behavioral immune system shapes political intuitions: Why and how individual differences in disgust sensitivity underlie opposition to immigration." In: *American Political Science Review* 111.2, pp. 277–294. ISSN: 15375943. DOI: `10.1017/S0003055416000770`.

Ahn, Woo Young et al. (2014). "Nonpolitical images evoke neural predictors of political ideology." In: *Current Biology* 24.22, pp. 2693–2699. ISSN: 09609822. DOI: `10.1016/j.cub.2014.09.050`. URL: `http://dx.doi.org/10.1016/j.cub.2014.09.050`.

Allen, Mark, William Badecker, and Lee Osterhout (2003). "Morphological analysis in sentence processing: An ERP study." In: *Language and Cognitive Processes* 18.4, pp. 405–430.

Altmann, Gerry T.M. and Yuki Kamide (1999). "Incremental interpretation at verbs: Restricting the domain of subsequent reference." In: *Cognition* 73.3, pp. 247–264.

Andrews, Mark, Gabriella Vigliocco, and David Vinson (2009). "Integrating Experiential and Distributional Data to Learn Semantic Representations." In: *Psychological Review* 116.3, pp. 463–498. ISSN: 0033295X. DOI: `10.1037/a0016261`.

Aronovitch, Charles D. (1976). "The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker." In: *Journal of Social Psychology* 99.2, pp. 207–220. ISSN: 19401183. DOI: `10.1080/00224545.1976.9924774`.

Arppe, Antti and Juhani Järvikivi (2007a). "Every method counts: Combining corpus-based and experimental evidence in the study of synonymy." In: *Corpus Linguistics and Linguistic Theory* 3.2, pp. 131–159.

Arppe, Antti and Juhani Järvikivi (2007b). "Take empiricism seriously! In support of methodological diversity in linguistics [Commentary of Geoffrey Sampson 2007: Grammar without Grammaticality.]" In: *Corpus Linguistics and Linguistic Theory* 3.1, pp. 99–109. ISSN: 16137027. DOI: `10.1515/CLLT.2006.007`.

Ashton, Michael C. and Kibeom Lee (2007). "Empirical, theoretical, and practical advantages of the HEXACO model of personality structure." In: *Personality and social psychology review* 11.2, pp. 150–166.

Aunger, Robert and Valerie Curtis (2013). "The anatomy of motivation: An evolutionary-ecological approach." In: *Biological Theory* 8.1, pp. 49–63.

Babel, Molly (2010). "Dialect divergence and convergence in New Zealand English." In: *Language in Society* 39.4, pp. 437–456. ISSN: 0047-4045. DOI: `10.1017/s0047404510000400`.

Banaji, Mahzarin R. and Curtis D. Hardin (1996). "Automatic Stereotyping." In: *Psychological Science* 7.3, pp. 136–141.

Bargh, John A. and Tanya L. Chartrand (1999). “The Unbearable Automaticity of Being.” In: *American Psychologist* 54.7, pp. 462–479.

Baron-Cohen, Simon, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb (2001). “The “Reading the Mind in the Eyes” Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism.” In: *The Journal of Child Psychology and Psychiatry and Allied Disciplines* 42.2, pp. 241–251.

Barsalou, Lawrence W. (2008). “Grounded Cognition.” In: *Annual Review of Psychology* 59.1, pp. 617–645. ISSN: 0066-4308. DOI: `10.1146/annurev.psych.59.103006.093639`.

Bartoń, Kamil (2018). *MuMIn: Multi-Model Inference.* R package version 1.42.1. URL: `https://CRAN.R-project.org/package=MuMIn`.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). “Fitting Linear Mixed-Effects Models Using lme4.” In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: `10.18637/jss.v067.i01`.

Baumeister, Roy E., Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice (1998). “Ego Depletion: Is the Active Self a Limited Resource?” In: *Journal of Personality and Social Psychology* 74.5, pp. 1252–1265.

Beatty, Jackson (1982). “Task-evoked pupillary responses, processing load, and the structure of processing resources.” In: *Psychological Bulletin* 91.2, pp. 276–292. ISSN: 00332909. DOI: `10.1037/0033-2909.91.2.276`. arXiv: `0112017 [cs]`.

Belin, Pascal, Patricia E.G. Bestelmeyer, Marianne Latinus, and Rebecca Watson (2011). “Understanding Voice Perception.” In: *British Journal of Psychology* 102.4, pp. 711–725. ISSN: 20448295. DOI: `10.1111/j.2044-8295.2011.02041.x`.

Bergamin, Oliver, Andreas Schoetzau, Keiko Sugimoto, and Mario Zulauf (1998). “The influence of iris color on the pupillary light reflex.” In: *Graefe’s archive for clinical and experimental ophthalmology* 236.8, pp. 567–570.

Boersma, Paul and David Weenink (2016). *Praat: doing phonetics by computer [Computer program].* Version 6.0.19, retrieved July 2016 from http://www.praat.org/.

Boland, Julie E. and Robin Queen (2016). “If You’re house is still available, send me an email: Personality influences reactions to written errors in email messages.” In: *PloS one* 11.3, e0149885.

Boroditsky, Lera, Michael Ramscar, and Michael C. Frank (2001). “The Roles of Body and Mind in Abstract Thought.” In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 23, pp. 276–281. DOI: `https://doi.org/ISBN978-0-9768318-8-4`.

Bradley, Margaret M., Laura Miccoli, Miguel A. Escrig, and Peter J. Lang (2008). “The pupil as a measure of emotional arousal and autonomic activation.” In: *Psychophysiology* 45.4, pp. 602–607. ISSN: 00485772. DOI: `10.1111/j.1469-8986.2008.00654.x`. arXiv: `NIHMS150003`.

Bradlow, Ann R., Gina M. Torretta, and David B. Pisoni (1996). “Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics.” In: *Speech Communication* 20.3-4, pp. 255–272. ISSN: 01676393. DOI: `10.1016/S0167-6393(96)00063-5`.

Braze, David, Donald Shankweiler, Weijia Ni, and Laura Conway Palumbo (2002). “Readers’ eye movements distinguish anomalies of form and content.” In: *Journal of psycholinguistic research* 31.1, pp. 25–44.

Busato, Vittorio V., Frans J. Prins, Jan J. Elshout, and Christiaan Hamaker (1998). "The relation between learning styles, the Big Five personality traits and achievement motivation in higher education." In: *Personality and individual differences* 26.1, pp. 129–140.

Busato, Vittorio V., Frans J. Prins, Jan J. Elshout, and Christiaan Hamaker (2000). "Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education." In: *Personality and Individual differences* 29.6, pp. 1057–1068.

Bybee, Joan L. (2010). *Language, Usage and Cognition.* Cambridge University Press. ISBN: 9780521851404. URL: `http://login.ezproxy.library.ualberta.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=e000xna&AN=320473&site=ehost-live&scope=site`.

Cantor, Nancy and Walter Mischel (1979). "Prototypes in person perception." In: *Advances in experimental social psychology.* Vol. 12. Elsevier, pp. 3–52.

Carreiras, Manuel, Alan Garnham, Jane Oakhill, and Kate Cain (1996). "The Use of Stereotypical Gender Information in Constructing a Mental Model: Evidence from English and Spanish." In: *The Quarterly Journal of Experimental Psychology* 49A.3, pp. 639–663. ISSN: 0959-2318. DOI: `10.1080/09592319908423242`.

Cawvey, Matthew, Matthew Hayes, Damarys Canache, and Jeffery J Mondak (2016). *Personality and Political Behavior.* DOI: `10.1093/acrefore/9780190228637.013.221`.

Chambers, Craig G., Michael K. Tanenhaus, and James S. Magnuson (2004). "Actions and affordances in syntactic ambiguity resolution." In: *Journal of experimental psychology: Learning, memory, and cognition* 30.3, p. 687.

Chamorro-Premuzic, Tomas and Adrian Furnham (2009). "Mainly Openness: The relationship between the Big Five personality traits and learning approaches." In: *Learning and Individual Differences* 19.4, pp. 524–529.

Chapman, Colin D., Christian Benedict, and Helgi B. Schiöth (2018). "Experimenter gender and replicability in science." In: *Science advances* 4.1, e1701427.

Chapman, Hanah A., Kristen Johannes, Jordan L Poppenk, Morris Moscovitch, and Adam K. Anderson (2013). "Evidence for the differential salience of disgust and fear in episodic memory." In: *Journal of Experimental Psychology: General* 142.4, p. 1100.

Chapman, Hanah A., David A. Kim, J. M. Susskind, and Adam K. Anderson (2009). "In bad taste: evidence for the oral origins of moral disgust." In: *Science* 323.5918, pp. 1222–1226. ISSN: 0036-8075. DOI: `10.1126/science.1165565`.

Chomsky, Noam (1957). *Syntactic Structures.* Mouton & Co.

Clahsen, Harald (2008). "Behavioral Methods for Investigating Morphological and Syntactic Processing in Children." In: *Developmental Psycholinguistics: On-Line Methods in Children's Language Processing.* Ed. by Irina A. Sekerina, Eva M. Fernandez, and Harald Clahsen. Amsterdam/Philadelphia: John Benjamins, pp. 1–28.

Clark, Andy (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science." In: *Behavioral and Brain Sciences* 36.3, pp. 181–204.

Coats, Susan and Eliot R. Smith (2007). "Perceptions of Gender Subtypes: Sensitivity to Recent Exemplar Activation and In-Group/Out-Group Differences." In: *Personality and Social Psychology Bulletin* 25.4, pp. 516–526. ISSN: 0146-1672. DOI: `10.1177/0146167299025004009`.

Cottrell, Catherine A. and Steven L. Neuberg (2005). “Different emotional reactions to different groups: a sociofunctional threat-based approach to” prejudice”.” In: *Journal of personality and social psychology* 88.5, p. 770.

Coulson, Seana, Jonathan W. King, and Marta Kutas (1998). “Expect the unexpected: Event-related brain response to morphosyntactic violations.” In: *Language and cognitive processes* 13.1, pp. 21–58.

Crippa, Sylvain Vincent, Fatima Pedrosa Domellöf, and Aki Kawasaki (2018). “Chromatic pupillometry in children.” In: *Frontiers in neurology* 9, p. 669.

Croft, William and D. Alan Cruse (2004). *Cognitive Linguistics.* DOI: `10.1192/bjp.112.483.211-a`.

Cutler, Anne and Charles Clifton (1999). “Comprehending spoken language: a blueprint of the listener.” In: *The neurocognition of language*, pp. 123–166.

da Silva-Castanheira, Kevin, Myles LoParco, and A. Ross Otto (2019). “Pupillometry as a Measure of Effort Exertion in Cognitive Control Tasks.” In: *Proceedings of the 25th annual conference of the cognitive science society.* Montreal, QC, p. 3439.

Davies, Mark (2008). *The Corpus of Contemporary American English: 520 million words, 1990-present.* Available online at http://corpus.byu.edu/coca/.

Davis, Mark H. (1980). “A multidimensional approach to individual differences in empathy.” In:

De Raad, Boele and Henri C. Schouwenburg (1996). “Personality in learning and education: A review.” In: *European Journal of personality* 10.5, pp. 303–336.

De Vincenzi, Marica, Remo Job, Rosalia Di Matteo, Alessandro Angrilli, Barbara Penolazzi, Laura Ciccarelli, and Francesco Vespignani (2003). “Differences in the perception and time course of syntactic and semantic violations.” In: *Brain and language* 85.2, pp. 280–296.

DeLong, Katherine A., Marta Kutas, and Thomas P. Urbach (2005). “Probabilistic word pre-activation during language comprehension inferred from electrical brain activity.” In: *Nature neuroscience* 8.8, p. 1117.

Dewaele, Jean-Marc and Adrian Furnham (1999). “Extraversion: The unloved variable in applied linguistic research.” In: *Language Learning* 49.3, pp. 509–544.

Dewaele, Jean-Marc and Adrian Furnham (2000). “Personality and speech production: a pilot study of second language learners.” In: *Personality and Individual differences* 28.2, pp. 355–365.

Ditman, Tali, Phillip J. Holcomb, and Gina R. Kuperberg (2007). “An investigation of concurrent ERP and self-paced reading methodologies.” In: *Psychophysiology* 44.6, pp. 927–935.

Divjak, Dagmar, Antti Arppe, and R. Harald Baayen (2016). “Does language-as-used fit a self-paced reading paradigm?” In: *Slavic Languages in Psycholinguistics* May, pp. 52–82.

Druschel, Barry A. and Martin F. Sherman (1999). “Disgust sensitivity as a function of the Big Five and gender.” In: *Personality and Individual Differences* 26.4, pp. 739–748. ISSN: 01918869. DOI: `10.1016/S0191-8869(98)00196-2`.

Eysenck, Hans J (1990). “Biological dimensions of personality.” In: *Handbook of personality: Theory and research.* Ed. by L. A. Pervin. New York: Guilford, pp. 244–276.

Faulkner, Jason, Mark Schaller, Justin H. Park, and Lesley A. Duncan (2004). "Evolved disease-avoidance mechanisms and contemporary xenophobic attitudes." In: *Group Processes and Intergroup Relations* 7.4, pp. 333–353. ISSN: 13684302. DOI: 10.1177/1368430204046142.

Federmeier, Kara D. (2007). "Thinking ahead: The role and roots of prediction in language comprehension." In: *Psychophysiology* 44.4, pp. 491–505.

Ferreira, Fernanda (2003). "The misinterpretation of noncanonical sentences." In: *Cognitive Psychology* 47.2, pp. 164–203. ISSN: 00100285. DOI: 10.1016/S0010-0285(03)00005-7.

Fox, John and Jangman Hong (2009). "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package." In: *Journal of Statistical Software* 32.1, pp. 1–24. URL: http://www.jstatsoft.org/v32/i01/.

Furnham, Adrian, David J. Hughes, and Emma Marshall (2013). "Creativity, OCD, Narcissism and the Big Five." In: *Thinking Skills and Creativity* 10, pp. 91–98. ISSN: 18711871. DOI: 10.1016/j.tsc.2013.05.003. URL: http://dx.doi.org/10.1016/j.tsc.2013.05.003.

Furnham, Adrian, Chris J. Jackson, and Tony Miller (1999). "Personality, learning style and work performance." In: *Personality and individual differences* 27.6, pp. 1113–1122.

Geeraerts, Dirk (2013). "Methodology in Cognitive Linguistics." In: *Cognitive linguistics: Current applications and future perspectives.* Ed. by Gitte Kristiansen. Psychology Press, pp. 21–49.

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling, and Shang E. Ha (2010). "Personality and political attitudes: Relationships across issue domains and political contexts." In: *American Political Science Review* 104.1, pp. 111–133. ISSN: 00030554. DOI: 10.1017/S0003055410000031.

Gibbs, Raymond W. (1994). *The poetics of mind: Figurative thought, language, and understanding.* Cambridge University Press.

Gibbs, Raymond W. and Marcus Perlman (2010). "Language understanding is grounded in experiential simulations: A response to Weiskopf." In: *Studies in History and Philosophy of Science* 41.3, pp. 305–308. ISSN: 00393681. DOI: 10.1016/j.shpsa.2010.07.004.

Gill, Alastair J. and Jon Oberlander (2002). "Taking care of the linguistic features of extraversion." In: *Proceedings of the Cognitive Science Society.* Vol. 24. 24.

Gill, Alastair J. and Jon Oberlander (2003). "Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry." In: *Proceedings of the 25th annual conference of the cognitive science society.* Erlbaum Hillsdale, NJ, pp. 456–461.

Gingras, Bruno, Manuela M. Marin, Estela Puig-Waldmüller, and W.T. Fitch (2015). "The eye is listening: Music-induced arousal and individual differences predict pupillary responses." In: *Frontiers in human neuroscience* 9.

Glenberg, Arthur M. and Michael P. Kaschak (2002). "Grounding language in action." In: *Psychonomic bulletin & review* 9.3, pp. 558–565.

Goldinger, Stephen D. and Megan H. Papesh (2012). "Pupil dilation reflects the creation and retrieval of memories." In: *Current Directions in Psychological Science* 21.2, pp. 90–95.

Goodman, Noah D. and Andreas Stuhlmüller (2013). "Knowledge and Implicature: Modeling Language Understanding as Social Cognition." In: *Topics in Cognitive Science* 5.1, pp. 173–184. DOI: 10.1111/tops.12007.

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek (2009). "Liberals and Conservatives Rely on Different Sets of Moral Foundations." In: *Journal of Personality and Social Psychology* 96.5, pp. 1029–1046. ISSN: 00223514. DOI: 10.1037/a0015141.

Grenier, Charles (n.d.). *Political Ideology Questionnaire*. School of Social Work, Louisiana State University. Retrieved from http://www.abacon.com/popplesw/ideology.pdf.

Grey, Sarah and Janet Van Hell (2017). "Foreign-accented speaker identity affects neural correlates of language comprehension." In: *Journal of Neurolinguistics* 42, pp. 93–108.

Grice, Paul (1975). "Logic and Conversation." In: *Syntax and Semantics*. Ed. by P Cole and J L Morgan. Vol. 3. New York: Seminar Press, pp. 41–58. DOI: 10.1111/j.1365-2664.2006.01229.x. eprint: arXiv:1011.1669v3.

Griffiths, Thomas L., Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum (2010). "Probabilistic models of cognition: Exploring representations and inductive biases." In: *Trends in cognitive sciences* 14.8, pp. 357–364.

Gruber, Kenneth J. and Jacquelyn Gaebelein (1979). "Sex differences in listening comprehension." In: *Sex Roles* 5.3, pp. 299–310. URL: https://libres.uncg.edu/ir/uncg/f/K%7B%5C_%7DGruber%7B%5C_%7DSex%7B%5C_%7D1979.pdf.

Gurven, Michael, Christopher von Rueden, Maxim Massenkoff, Hillard Kaplan, and Marino Lero Vie (2013). "How Universal Is the Big Five? Testing the Five-Factor Model of Personality Variation Among Forager–Farmers in the Bolivian Amazon." In: *Journal of personality and social psychology* 104.2, pp. 354–370. DOI: doi:10.1037/a0030841.

Haapalainen, Eija, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey (2010). "Psychophysiological measures for assessing cognitive load." In: *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, pp. 301–310.

Hagoort, Peter, Lea Hald, Marcel Bastiaansen, and Karl M. Petersson (2004). "Integration of Word Meaning and World Knowledge in Language Comprehension." In: *Science* 304.5669, pp. 438–441. ISSN: 0036-8075. DOI: 10.1126/science.1095455. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1095455.

Hagoort, Peter and Jos J.A. Van Berkum (2007). "Beyond the sentence given." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481, pp. 801–811. ISSN: 09628436. DOI: 10.1098/rstb.2007.2089.

Haidt, McCauley & Rozin, 1994, modified by Olatunji et al. (2007). *The DS-R*.

Haidt, Jonathan and Jesse Graham (2007). "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize." In: *Social Justice Research* 20.1, pp. 98–116. ISSN: 08857466. DOI: 10.1007/s11211-007-0034-z. arXiv: arXiv:1011.1669v3.

Haidt, Jonathan, Clark McCauley, and Paul Rozin (1994). "Individual-Differences in Sensitivity To Disgust - a Scale Sampling 7 Domains of Disgust Elicitors." In: *Personality and Individual Differences* 16.5, pp. 701–713. DOI: 10.1016/0191-8869(94)90212-7.

Hanulíková, Adriana and Manuel Carreiras (2015). "Electrophysiology of subject-verb agreement mediated by speakers' gender." In: *Frontiers in Psychology* 6.September. ISSN: 16641078. DOI: 10.3389/fpsyg.2015.01396.

Hanulíková, Adriana, Petra M. Van Alphen, Merel M. Van Goch, and Andrea Weber (2012). "When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing." In: *Journal of Cognitive Neuroscience* 24.4, pp. 878–887.

Harris, Catherine L., Ayşe Ayçiçeği, and Jean Berko Gleason (2003). "Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language." In: *Applied Psycholinguistics* 24.4, pp. 561–579.

Haspelmath, Martin (2002). "Functionalist linguistics: usage-based explanations of language structure." In: *Handout from talk given at Düsseldorf Summer School*, pp. 1–64. URL: `http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:Functionalist+linguistics:+usage-based+explanations+of+language+structure%7B%5C#%7D0`.

Hatemi, Peter K. and Brad Verhulst (2015). "Political attitudes develop independently of personality traits." In: *PLoS ONE* 10.3, pp. 1–25. ISSN: 19326203. DOI: `10.1371/journal.pone.0118106`.

Havas, David A., Arthur M. Glenberg, Karol A. Gutowski, Mark J. Lucarelli, and Richard J. Davidson (2010). "Cosmetic use of botulinum toxin-a affects processing of emotional language." In: *Psychological Science* 21.7, pp. 895–900. ISSN: 09567976. DOI: `10.1177/0956797610374742`.

Havas, David A., Arthur M. Glenberg, and Mike Rinck (2007). "Emotion simulation during language comprehension." In: *Psychonomic Bulletin & Review* 14.3, pp. 436–441.

Heinström, Jannica (2005). "Fast surfing, broad scanning and deep diving: The influence of personality and study approach on students' information-seeking behavior." In: *Journal of documentation* 61.2, pp. 228–247.

Helzer, Erik G. and David A. Pizarro (2011). "Dirty liberals!: Reminders of physical cleanliness influence moral and political attitudes." In: *Psychological Science* 22.4, pp. 517–522. ISSN: 09567976. DOI: `10.1177/0956797611402514`.

Hilton, James L. and William von Hippel (1996). "Stereotypes." In: *Annual review of psychology* 47, pp. 237–271.

Hinton, Geoffrey E. (2007). "Learning multiple layers of representation." In: *Trends in cognitive sciences* 11.10, pp. 428–434.

Hlavac, Marek (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables.* R package version 5.2.2. Central European Labour Studies Institute (CELSI). Bratislava, Slovakia. URL: `https://CRAN.R-project.org/package=stargazer`.

Huettig, Falk (2015). "Four central questions about prediction in language processing." In: *Brain Research* 1626, pp. 118–135.

Huettig, Falk and Nivedita Mani (2016). "Is prediction necessary to understand language? Probably not." In: *Language, Cognition and Neuroscience* 31.1, pp. 19–31.

Ibbotson, Paul (2013). "The scope of usage-based theory." In: *Frontiers in Psychology* 4.MAY, pp. 1–15. ISSN: 16641078. DOI: `10.3389/fpsyg.2013.00255`.

Inbar, Yoel, David A. Pizarro, and Paul Bloom (2009). "Conservatives are more easily disgusted than liberals." In: *Cognition and Emotion* 23.4, pp. 714–725. ISSN: 02699931. DOI: `10.1080/02699930802110007`.

Inbar, Yoel, David A Pizarro, Ravi Iyer, and Jonathan Haidt (2011). "Disgust Sensitivity, Political Conservatism, and Voting." In: *Social Psychological and Personality Science.* ISSN: 10000569. DOI: `10.1177/1948550611429024`.

Jackson, Philip L. and Jean Decety (2004). "Motor cognition: A new paradigm to study self–other interactions." In: *Current opinion in neurobiology* 14.2, pp. 259–263.

Jackson, Philip L, Eric Brunet, Andrew N. Meltzoff, and Jean Decety (2006). “Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain.” In: *Neuropsychologia* 44.5, pp. 752–761.

Jang, Kwang Ming (1998). “Eysenck’s PEN model: Its Contribution to personality Psychology.” In: *Contribution of Eysenck’s PEN Model.*

Jegerski, Jill and Bill VanPatten, eds. (2013). *Research Methods in Second Language Psycholinguistics*. Second Language Acquisition Research Series. New York: Routledge. URL: `https://books.google.ca/books?id=tMhiAgAAQBAJ`.

Jensen, Mikael (2015). “Personality traits, learning and academic achievements.” In: *Journal of Education and Learning* 4.4, p. 91.

John, Oliver P. and Sanjay Srivastava (1999). “The Big Five trait taxonomy: History, measurement, and theoretical perspectives.” In: *Handbook of personality: Theory and research* 2.1999, pp. 102–138.

John, Oliver P, E. M. Donahue, and Kentle. R. L. (1991). *The Big Five Inventory–Versions 4a and 54*. University of California, Berkeley, Institute of Personality and Social Research. DOI: `10.1016/S0092-6566(03)00046-1`. URL: `http://linkinghub.elsevier.com/retrieve/pii/S0092656603000461`.

Johnston, Lucy and Miles Hewstone (1992). “Cognitive models of stereotype change: 3. Subtyping and the perceived typicality of disconfirming group members.” In: *Journal of Experimental Social Psychology* 28.4, pp. 360–386.

Jost, John T., Jack Glaser, Arie W. Kruglanski, and Frank J Sulloway (2003). “Political Conservatism as Motivated Social Cognition John.” In: *Psychological Bulletin* 129.3, pp. 339–375. ISSN: 00113891. DOI: `10.1037/0033-2909.129.3.339`.

Just, Marcel A. and Patricia A. Carpenter (1980). “A theory of reading: From eye fixations to comprehension.” In: *Psychological review* 87.4, p. 329.

Just, Marcel A. and Patricia A. Carpenter (1993). “The intensity dimension of thought: pupillometric indices of sentence processing.” In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 47.2, p. 310.

Kahneman, Daniel and Jackson Beatty (1966). “Pupil diameter and load on memory.” In: *Science* 154.3756, pp. 1583–1585.

Kamide, Yuki, Gerry T.M. Altmann, and Sarah L. Haywood (2003). “The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements.” In: *Journal of Memory and language* 49.1, pp. 133–156.

Katzir, Maayan, Matan Hoffmann, and Nira Liberman (2018). “Disgust as an Essentialist Emotion That Signals Nonviolent Outgrouping With Potentially Low Social Costs.” In: *Emotion* 19.5, pp. 841–862. ISSN: 19311516. DOI: `10.1037/emo0000480`.

Kaup, Barbara, Jana Lüdtke, and Rolf A. Zwaan (2007). “The experiential view of language comprehension: How is negation represented?” In: *Higher level language processes in the brain: Inference and comprehension processes*. Ed. by Franz Schmalhofer and C. A Perfetti. London: Lawrence Erlbaum Associates, pp. 255–288. ISBN: 978-0-8058-5262-2—0-8058-5262-X.

Klein, Stanley B., Judith Loftus, and John F. Kihlstrom (2002). “Memory and temporal experience: The effects of episodic memory loss on an amnesic patient’s ability to remember the past and imagine the future.” In: *Social Cognition* 20.5, pp. 353–379. ISSN: 0278016X. DOI: `10.1521/soco.20.5.353.21125`.

Knill, David C. and Alexandre Pouget (2004). “The Bayesian brain: the role of uncertainty in neural coding and computation.” In: *TRENDS in Neurosciences* 27.12, pp. 712–719.

Knoeferle, Pia, Matthew W. Crocker, Christoph Scheepers, and Martin J. Pickering (2005). “The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events.” In: *Cognition* 95.1, pp. 95–127.

Ko, Sei Jin, Charles M. Judd, and Irene V. Blair (2006). “What the voice reveals: Within- and between-category stereotyping on the basis of voice.” In: *Personality and Social Psychology Bulletin* 32.6, pp. 806–819. ISSN: 01461672. DOI: 10.1177/0146167206286627.

Komarraju, Meera and Steven J. Karau (2005). “The relationship between the big five personality traits and academic motivation.” In: *Personality and individual differences* 39.3, pp. 557–567.

Kristiansen, Gitte (2006). *Cognitive linguistics: Current applications and future perspectives.* ISBN: 9783110189506.

Kutas, Marta (1997). “Views on how the electrical activity that the brain generates reflects the functions of different language structures.” In: *Psychophysiology* 34.4, pp. 383–398.

Kutas, Marta and Kara D. Federmeier (2007). “Event-Related brain potential (ERP) studies of sentence processing.” In: *Oxford Handbook of Psycholinguistics.* Chap. 23, pp. 385–406. DOI: 10.1093/oxfordhb/9780198568971.013.0023.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2017). “lmerTest Package: Tests in Linear Mixed Effects Models.” In: *Journal of Statistical Software* 82.13, pp. 1–26. DOI: 10.18637/jss.v082.i13.

Lakoff, George and Mark Johnson (1980). “The metaphorical structure of the human conceptual system.” In: *Cognitive science* 4.2, pp. 195–208.

Lakoff, George and Mark Johnson (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought.* Vol. 4. New york: Basic books.

Lattner, Sonja and Angela D. Friederici (2003). “Talker’s voice and gender stereotype in human auditory sentence processing - Evidence from event-related brain potentials.” In: *Neuroscience Letters* 339, pp. 191–194. ISSN: 03043940. DOI: 10.1016/S0304-3940(03)00027-2.

Lee, Kibeom and Michael C. Ashton (2004). “Psychometric properties of the HEXACO personality inventory.” In: *Multivariate behavioral research* 39.2, pp. 329–358.

Levy, Roger (2008). “Expectation-based syntactic comprehension.” In: *Cognition* 106.3, pp. 1126–1177.

Levy, Roger (2010). “A noisy-channel model of rational human sentence comprehension under uncertain input.” In: October, p. 234. DOI: 10.3115/1613715.1613749.

Li, Sai, Xiaoming Jiang, Hongbo Yu, and Xiaolin Zhou (2014). “Cognitive empathy modulates the processing of pragmatic constraints during sentence comprehension.” In: *Social cognitive and affective neuroscience* 9.8, pp. 1166–1174.

Linville, Patricia W., Gregory W. Fischer, and Peter Salovey (1989). “Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation.” In: *Journal of personality and social psychology* 57.2, p. 165.

Lõo, Kaidi, Jacolien van Rij, Juhani Järvikivi, and Harald Baayen (2016). “Individual Differences in Pupil Dilation during Naming Task.” In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society, A. Papafragou, D. Grodner, D. Mirman, and J. Trueswell, Eds. Austin, TX: Cognitive Science Society*, pp. 550–555.

Macrae, C. Neil and Galen V. Bodenhausen (2000). “Thinking Categorically about Others.” In: *Annual review of psychology* 51, pp. 93–120.

Marinis, Theodore (2003). “Psycholinguistic techniques in second language acquisition research.” In: *Second language research* 19.2, pp. 144–161.

Marrville, Caelan (2017). “Gender and dominance in action: World view and emotional affect in language processing and use.” PhD thesis. University of Alberta.

Matlock, Teenie, Michael Ramscar, and Lera Boroditsky (2005). “On the experiential link between spatial and temporal language.” In: *Cognitive Science* 29.4, pp. 655–664. ISSN: 03640213. DOI: 10.1207/s15516709cog0000_17.

McGinn, Colin (2011). *The Meaning of Disgust.* Oxford University Press. DOI: 10.1093/acprof:oso/9780199829538.001.0001.

McRae, Ken and Kazunaga Matsuki (2013). “Constraint-based Models of Sentence Processing.” In: *Current Issues in the Psychology of Language. Sentence Processing.* Ed. by Roger P. G. van Gompel. New York, NY: Psychology Press, pp. 51–77.

Michel, Jean-Baptiste et al. (2011). “Quantitative analysis of culture using millions of digitized books.” In: *Science* 331.6014, pp. 176–182.

Molinaro, Nicola, Jui Ju Su, and Manuel Carreiras (2016). “Stereotypes override grammar: Social knowledge in sentence comprehension.” In: *Brain and Language* 155-156, pp. 36–43. ISSN: 10902155. DOI: 10.1016/j.bandl.2016.03.002. URL: http://dx.doi.org/10.1016/j.bandl.2016.03.002.

Moore, Kelly and James C. McElroy (2012). “The influence of personality on Facebook usage, wall postings, and regret.” In: *Computers in Human Behavior* 28.1, pp. 267–274.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman (2012). “Science faculty’s subtle gender biases favor male students.” In: *Proceedings of the National Academy of Sciences* 109.41, pp. 16474–16479. ISSN: 0027-8424. DOI: 10.1073/pnas.1211286109. arXiv: arXiv:1408.1149. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1211286109.

Mukai, Yoichi, Juhani Järvikivi, and Benjamin V. Tucker (2018). “The effect of phonological-orthographic consistency on the processing of reduced and citation forms of Japanese words: Evidence from pupillometry.” In: *The Annual Conference of the Canadian Linguistic Association.*

Murray, Damian R. and Mark Schaller (2016). “The behavioral immune system: Implications for social cognition, social interaction, and social influence.” In: *Advances in Experimental Social Psychology* 53, pp. 75–129. ISSN: 00652601. DOI: 10.1016/bs.aesp.2015.09.002.

Neuberg, Steven L., Douglas T. Kenrick, and Mark Schaller (2011). “Human threat management systems: Self-protection and disease avoidance.” In: *Neuroscience and Biobehavioral Reviews* 35.4, pp. 1042–1051. ISSN: 01497634. DOI: 10.1016/j.neubiorev.2010.08.011. arXiv: NIHMS150003.

Ni, Weijia, Janet Dean Fodor, Stephen Crain, and Donald Shankweiler (1998). “Anomaly detection: Eye movement patterns.” In: *Journal of Psycholinguistic Research* 27.5, pp. 515–539.

Nieuwland, Mante S. and Jos J.A. Van Berkum (2006). “When peanuts fall in love: N400 evidence for the power of discourse.” In: *Journal of cognitive neuroscience* 18.7, pp. 1098–1111.

Oakhill, Jane, Alan Garnham, and David Reynolds (2005). “Immediate activation of stereotypical gender information.” In: *Memory & Cognition* 33.6, pp. 972–983.

Oaten, Megan, Richard J. Stevenson, and Trevor I. Case (2009). “Disgust as a disease-avoidance mechanism.” In: *Psychological bulletin* 135.2, p. 303.

Oaten, Megan, Richard J. Stevenson, and Trevor I. Case (2011). “Disease avoidance as a functional basis for stigmatization.” In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1583, pp. 3433–3452.

Oberlander, Jon and Alastair J. Gill (2004). “Individual differences and implicit language: personality, parts-of-speech and pervasiveness.” In: *Proceedings of the Cognitive Science Society*. Vol. 26. 26.

Orlob, Chris (2017). *Men Listen to Women Less Than They Do to Other Men, Data Suggests*. LinkedIn Pulse. Retrieved from https://www.linkedin.com/pulse/men-listen-women-less-than-do-other-datasuggests-chris-orlob/.

Ormel, Johan et al. (2013). “Neuroscience and Biobehavioral Reviews The biological and psychological basis of neuroticism : Current status and future directions.” In: *Neuroscience and Biobehavioral Reviews* 37.1, pp. 59–72. ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2012.09.004. URL: http://dx.doi.org/10.1016/j.neubiorev.2012.09.004.

Osterhout, Lee (1999). “A superficial resemblance does not necessarily mean you are part of the family: Counterarguments to Coulson, King and Kutas (1998) in the P600/SPS-P300 debate.” In: *Language and Cognitive Processes* 14.1, pp. 1–14.

Osterhout, Lee, Michael Bersick, and Judith McLaughlin (1997). “Brain potentials reflect violations of gender stereotypes.” In: *Memory and Cognition* 25.3, pp. 273–285. ISSN: 0090502X. DOI: 10.3758/BF03211283.

Otten, Marte and Jos J.A. Van Berkum (2008). “Discourse-based word anticipation during language processing: Prediction or priming?” In: *Discourse Processes* 45.6, pp. 464–496.

Papadopoulou, Despina, Ianthi Tsimpli, and Nikos Amvrazis (2013). “Self-Paced Listening.” In: *Research Methods in Second Language Psycholinguistics*. Ed. by Jill Jegerski and Bill VanPatten. New York: Routledge, pp. 50–68.

Papesh, Megan H. and Stephen D. Goldinger (2012). “Pupil-BLAH-metry: Cognitive effort in speech planning reflected by pupil dilation.” In: *Attention, Perception, & Psychophysics* 74.4, pp. 754–765.

Park, Bernadette, Charles M. Judd, and Carey S. Ryan (1991). “Social categorization and the representation of variability information.” In: *European review of social psychology* 2.1, pp. 211–245.

Park, Gregory et al. (2015). “Automatic personality assessment through social media language.” In: *Journal of personality and social psychology* 108.6, p. 934.

Parks-Leduc, Laura, Gilad Feldman, and Anat Bardi (2015). “Personality Traits and Personal Values: A Meta-Analysis.” In: *Personality and Social Psychology Review* 19.1, pp. 3–29. ISSN: 10888683. DOI: 10.1177/1088868314538548.

Partala, Timo and Veikko Surakka (2003). “Pupil size variation as an indication of affective processing.” In: *International Journal of Human Computer Studies* 59.1-2, pp. 185–198. ISSN: 10715819. DOI: 10.1016/S1071-5819(03)00017-X. arXiv: NIHMS150003.

Penke, Martina and Anette Rosenbach (2004). “What counts as evidence in linguistics?: An introduction.” In: *Studies in Language. International Journal sponsored by the Foundation “Foundations of Language”* 28.3, pp. 480–526.

Pennebaker, James W. and Laura A. King (1999). "Linguistic styles: language use as an individual difference." In: *Journal of personality and social psychology* 77.6, p. 1296.

Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer (2003). "Psychological aspects of natural language use: Our words, our selves." In: *Annual review of psychology* 54.1, pp. 547–577.

Pezzulo, Giovanni, Joachim Hoffmann, and Rino Falcone (2007). "Anticipation and anticipatory behavior." In: *Cognitive Processing* 8, pp. 67–70. DOI: 10.1007/s10339-007-0177-8.

Piquado, Tepring, Derek Isaacowitz, and Arthur Wingfield (2010). "Pupillometry as a measure of cognitive effort in younger and older adults." In: *Psychophysiology* 47.3, pp. 560–569.

Porretta, Vincent, Antoine Tremblay, and Patrick Bolger (2017). "Got experience? PMN amplitudes to foreign-accented speech modulated by listener experience." In: *Journal of Neurolinguistics* 44, pp. 54–67.

Porretta, Vincent and Benjamin V. Tucker (2019). "Eyes Wide Open: Pupillary Response to a Foreign Accent Varying in Intelligibility." In: *Frontiers in Communication* 4.February, pp. 1–12. DOI: 10.3389/fcomm.2019.00008.

Porretta, Vincent, Benjamin V. Tucker, and Juhani Järvikivi (2016). "The influence of gradient foreign accentedness and listener experience on word recognition." In: *Journal of Phonetics* 58, pp. 1–21.

Potter, Jonathan (2002). "The Malleability of Automatic Stereotypes and Prejudice." In: *Personality and Social Psychology Review*, pp. 192–194. ISSN: 1088-8683. DOI: 10.1207/S15327957PSPR0603.

Psychology Software Tools Inc. (2012). *E-Prime 2.0.* Retrieved from http://www.pstnet.com.

Pyykkönen, Pirita, Jukka Hyönä, and Roger P.G. Van Gompel (2010). "Activating gender stereotypes during online spoken language processing: Evidence from visual world eye tracking." In: *Experimental Psychology* 57.2, pp. 126–133. ISSN: 16183169. DOI: 10.1027/1618-3169/a000016.

Quadflieg, Susanne and C. Neil Macrae (2011). "Stereotypes and stereotyping: What's the brain got to do with it?" In: *European Review of Social Psychology* 22.1, pp. 215–273. ISSN: 1046-3283. DOI: 10.1080/10463283.2011.627998.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rammstedt, Beatrice (2007). "The 10-ltem big five inventory: Norm values and investigation of sociodemographic effects based on a German population representative sample." In: *European Journal of Psychological Assessment* 23.3, pp. 193–201. ISSN: 10155759. DOI: 10.1027/1015-5759.23.3.193.

Riegler, Alexander (2001). "The Role of Anticipation in Cognition." In: *Computing Anticipatory Systems. Proceedings of the American Institute of Physics 573.* Ed. by D M Dubois, pp. 534–541. DOI: 10.1063/1.1503666.

Rij, Jacolien van, Petra Hendriks, Hedderik van Rijn, R. Harald Baayen, and Simon N. Wood (2019). "Analyzing the Time Course of Pupillometric Data." In: *Trends in Hearing* 23, p. 233121651983248. ISSN: 2331-2165. DOI: 10.1177/2331216519832483. URL: http://journals.sagepub.com/doi/10.1177/2331216519832483.

Roberts, Leah (2012). “Psycholinguistic techniques and resources in second language acquisition research.” In: *Second Language Research* 28.1, pp. 113–127.

Robinson, David, Norman Gabriel, and Olga Katchan (1994). “Personality and second language learning.” In: *Personality and Individual Differences* 16.1, pp. 143–157.

Romero-Rivas, Carlos, Clara D. Martin, and Albert Costa (2016). “Foreign-accented speech modulates linguistic anticipatory processes.” In: *Neuropsychologia* 85, pp. 245–255.

Rondeel, Eefje W.M., Henk Van Steenbergen, Rob W. Holland, and Ad van Knippenberg (2015). “A closer look at cognitive control: differences in resource allocation during updating, inhibition and switching as revealed by pupillometry.” In: *Frontiers in human neuroscience* 9.

Rosenthal, Robert (1976). “Experimenter effects in behavioral research.” In:

Rubin, Donald L. (1992). “Nonlanguage factors affecting undergraduates’ judgments of nonnative English-speaking teaching assistants.” In: *Research in Higher Education* 33.4, pp. 511–531.

Sánchez-Bernardos, M. L., M. J. Hernández Lloreda, M. D. Avia, and C. Bragado-Alvarez (2015). “Fantasy Proneness and Personality Profiles.” In: *Imagination, Cognition and Personality* 34.4, pp. 327–339. ISSN: 0276-2366. DOI: 10.1177/0276236615572584.

Schacter, Daniel L., Donna Rose Addis, Demis Hassabis, Victoria C. Martin, R. Nathan Spreng, and Karl K. Szpunar (2012). “The Future of Memory: Remembering, Imagining, and the Brain.” In: *Neuron* 76.4, pp. 677–694. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.11.001. URL: http://dx.doi.org/10.1016/j.neuron.2012.11.001.

Schaller, Mark and Steven L. Neuberg (2012). *Danger, Disease, and the Nature of Prejudice(s).* 1st ed. Vol. 46. Elsevier Inc., pp. 1–54. DOI: 10.1016/B978-0-12-394281-4.00001-5. URL: http://dx.doi.org/10.1016/B978-0-12-394281-4.00001-5.

Schaller, Mark and Justin H. Park (2011). “The behavioral immune system (and why it matters).” In: *Current Directions in Psychological Science* 20.2, pp. 99–103. ISSN: 09637214. DOI: 10.1177/0963721411402596.

Schmidtke, Jens (2018). “Pupillometry in linguistic research: An introduction and review for second language researchers.” In: *Studies in Second Language Acquisition* 40.3, pp. 529–549.

Schmitt, David P. and Todd K. Shackelford (2008). “Big Five Traits Related to Short-Term Mating: From Personality to Promiscuity across 46 Nations.” In: *Evolutionary Psychology* 6.2, p. 147470490800600. ISSN: 1474-7049. DOI: 10.1177/147470490800600204. URL: http://journals.sagepub.com/doi/10.1177/147470490800600204.

Sedivy, Julie C., Michael K. Tanenhaus, Craig G. Chambers, and Gregory N. Carlson (1999). “Achieving incremental semantic interpretation through contextual representation.” In: *Cognition* 71.2, pp. 109–147.

Al-Shawaf, Laith, David M.G. Lewis, and David M. Buss (2018). “Sex Differences in Disgust: Why Are Women More Easily Disgusted Than Men?” In: *Emotion Review* 10.2, pp. 149–160. ISSN: 17540739. DOI: 10.1177/1754073917709940.

Sheldrake, Rupert (1998). “Experimenter effects in scientific research: How widely are they neglected.” In: *Journal of Scientific Exploration* 12.1, pp. 73–78.

Smith, Eliot R. and Michael A. Zárate (1992). “Exemplar-Based Model of Social Judgment.” In: *Psychological Review* 99.1, pp. 3–21. ISSN: 0033295X. DOI: 10.1037/0033-295X.99.1.3.

Smith, Kevin B., Douglas Oxley, Matthew V. Hibbing, John R. Alford, and John R. Hibbing (2011). "Disgust Sensitivity and the Neurophysiology of Left- Right Political Orientations." In: *PLoS ONE* 6.10. ISSN: 08846812. DOI: 10.1371/journal.pone.0025552.

Sparks, Adam M., Daniel M. T. Fessler, Kai Q. Chan, Ashwini Ashokkumar, and Colin Holbrook (2018). "Disgust as a mechanism of decision making under risk." In: *Emotion* 18.7, pp. 942–958.

Spivey, Michael J., Michael K. Tanenhaus, Kathleen M. Eberhard, and Julie C. Sedivy (2002). "Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution." In: *Cognitive psychology* 45.4, pp. 447–481.

Srivastava, Sanjay, Oliver P. John, Samuel D. Gosling, and Jeff Potter (2003). "Comparison Sample: Means and Standard Deviations for Big Five Inventory by Age." In: *Journal of Personality and Social Psychology* 84.5, p. 2003.

Steinhauer, Stuart R., Greg J. Siegle, Ruth Condray, and Misha Pless (2004). "Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing." In: *International Journal of Psychophysiology* 52.1, pp. 77–86. ISSN: 01678760. DOI: 10.1016/j.ijpsycho.2003.12.005.

Stenberg, Georg, Jarl Risberg, Siegbert Warkentin, and Ingmar Rosén (1990). "Regional patterns of cortical blood flow distinguish extraverts from introverts." In: *Personality and Individual Differences* 11.7, pp. 663–673.

Strand, Elizabeth A. (1999). "Uncovering the role of gender stereotypes in speech perception." In: *Journal of Language and Social Psychology* 18.1, pp. 86–100. ISSN: 0261927X. DOI: 10.1177/0261927X99018001006.

Strohminger, Nina (2014). "The meaning of disgust: A refutation." In: *Emotion Review* 6.3, pp. 214–216. ISSN: 17540739. DOI: 10.1177/1754073914523072.

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy (1995). "Integration of visual and linguistic information in spoken language comprehension." In: *Science*, pp. 1632–1634.

Tannen, Deborah (1993). "Marked Women, Unmarked Men." In: *The New York Times Magazine*, pp. 18–20.

Tokowicz, Natasha and Tessa Warren (2010). "Beginning adult L2 learners' sensitivity to morphosyntactic violations: A self-paced reading study." In: *European Journal of Cognitive Psychology* 22.7, pp. 1092–1106.

Trask, R. L. (1999). *Key Concepts in Language and Linguistics*. Vol. 76. 4, p. 949. ISBN: 0415157412. DOI: 10.2307/417239.

Traxler, Matthew J. (2014). "Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing." In: *Trends in Cognitive Sciences* 18.11, pp. 605–611. ISSN: 1879307X. DOI: 10.1016/j.tics.2014.08.001. URL: http://dx.doi.org/10.1016/j.tics.2014.08.001.

Tremblay, Antoine and Aaron J. Newman (2015). "Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples." In: *Psychophysiology* 52.1, pp. 124–139.

Tressoldi, Patrizio (2015). "Anticipation of Random Future Events." In: *Cognitive Systems Monographs: Anticipation across Disciplines* 29, pp. 1–403. DOI: 10.1007/978-3-319-22599-9.

Trueswell, John C., Michael K. Tanenhaus, and Susan M. Garnsey (1994). “Semantic Influences on Parsing: Use of Thematic Role Information in Syntactic Ambiguity Resolution.” In: *Journal of Memory and Language* 33, pp. 285–318.

Tybur, Joshua M., Debra Lieberman, Robert Kurzban, and Peter DeScioli (2013). “Disgust: Evolved function and structure.” In: *Psychological Review* 120.1, pp. 65–84. ISSN: 0033295X. DOI: 10.1037/a0030778.

Tybur, Joshua M. and Reinout E. de Vries (2013). “Disgust sensitivity and the HEXACO model of personality.” In: *Personality and Individual Differences* 55.6, pp. 660–665. ISSN: 01918869. DOI: 10.1016/j.paid.2013.05.008. URL: http://dx.doi.org/10.1016/j.paid.2013.05.008.

Uhlmann, Eric Luis and Geoffrey L. Cohen (2005). “Constructed Criteria: Redefining Merit to Justify Discrimination.” In: *Psychological Science* 16.6, pp. 474–480. ISSN: 0956-7976. DOI: 10.1111/j.0956-7976.2005.01559.x.

Van Berkum, Jos J.A., Colin M. Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort (2005). “Anticipating upcoming words in discourse: evidence from ERPs and reading times.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31.3, p. 443.

Van Berkum, Jos J.A., Dieuwke De Goede, Petra M. Van Alphen, Emma R. Mulder, and José H. Kerstholt (2013). “How robust is the language architecture? The case of mood.” In: *Frontiers in psychology* 4.

Van Berkum, Jos J.A., Bregje Holleman, Mante Nieuwland, Marte Otten, and Jaap Murre (2009). “Right or wrong? The brain’s fast response to morally objectionable statements.” In: *Psychological Science* 20.9, pp. 1092–1099.

Van Berkum, Jos J.A., Danielle Van den Brink, Cathelijne M.J.Y. Tesink, Miriam Kos, and Peter Hagoort (2008). “The neural integration of speaker and message.” In: *Journal of cognitive neuroscience* 20.4, pp. 580–591.

Van Boxtel, Geert and Koen Bocker (2004). “Cortical Measures of Anticipation.” In: *Journal of Psychophysiology* 18, pp. 61–76. ISSN: 0269-8803. DOI: 10.1027/0269-8803.18.2.

Van den Brink, Daniëlle, Jos J.A. Van Berkum, Marcel C.M. Bastiaansen, Cathelijne M.J.Y. Tesink, Miriam Kos, Jan K. Buitelaar, and Peter Hagoort (2010). “Empathy matters: ERP evidence for inter-individual differences in social language processing.” In: *Social cognitive and affective neuroscience* 7.2, pp. 173–183.

Van Petten, Cyma and Barbara J. Luka (2012). “Prediction during language comprehension: Benefits, costs, and ERP components.” In: *International Journal of Psychophysiology* 83.2, pp. 176–190.

van Rij, Jacolien, Martijn Wieling, R. Harald Baayen, and Hedderik van Rijn (2016). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.* R package version 2.2.

Vanpaemel, Wolf and Gert Storms (2008). “In search of abstraction: The varying abstraction model of categorization.” In: *Psychonomic bulletin & review* 15.4, pp. 732–749.

Vanpaemel, Wolf and Gert Storms (2010). “Abstraction and model evaluation in category learning.” In: *Behavior Research Methods* 42.2, pp. 421–437.

Verhulst, Brad, Lindon J. Eaves, and Peter K. Hatemi (2013). “Correlation not Causation: The Relationship between Personality Traits and Political Ideologies.” In: 6.8, pp. 34–51. ISSN: 1053-8119. DOI: 10.1021/nn300902w.Release. arXiv: NIHMS150003.

Viebahn, Malte C., Mirjam Ernestus, and James M. McQueen (2017). "Speaking style influences the brain's electrophysiological response to grammatical errors in speech comprehension." In: *Journal of Cognitive Neuroscience.*

Vogelzang, Margreet, Petra Hendriks, and Hedderik van Rijn (2016). "Pupillary responses reflect ambiguity resolution in pronoun processing." In: *Language, Cognition and Neuroscience* 3798, pp. 1–10. ISSN: 23273801. DOI: 10.1080/23273798.2016.1155718.

Webster, Steven W. (2018). "It's Personal: The Big Five Personality Traits and Negative Partisan Affect in Polarized U.S. Politics." In: *American Behavioral Scientist* 62.1, pp. 127–145. ISSN: 15523381. DOI: 10.1177/0002764218756925.

Wehrli, Stefan (2008). "Personality on social network sites: An application of the five factor model." In: *Zurich: ETH Sociology (Working Paper No. 7).*

Weibel, David, Corinna S. Martarelli, Diego Häberli, and Fred W. Mast (2018). "The Fantasy Questionnaire: A Measure to Assess Creative and Imaginative Fantasy." In: *Journal of Personality Assessment* 100.4, pp. 431–443. ISSN: 00223891. DOI: 10.1080/00223891.2017.1331913.

White, Andrew Edward, Douglas T. Kenrick, and Steven L. Neuberg (2013). "Beauty at the ballot box: Disease threats predict preferences for physically attractive leaders." In: *Psychological science* 24.12, pp. 2429–2436.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: http://ggplot2.org.

Wiggers, Kyle J. (2018). "Mei uses AI to improve relationships by analyzing text messages." In: *Venturebeat.*

Wilson, Glenn D. and John R. Patterson (1968). "A new measure of conservatism." In: *British Journal of Social and Clinical Psychology* 7.4, pp. 264–269.

Winn, Matthew B., Dorothea Wendt, Thomas Koelewijn, and Stefanie E. Kuchinsky (2018). "Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started." In: *Trends in Hearing* 22, p. 233121651880086. ISSN: 2331-2165. DOI: 10.1177/2331216518800869. URL: http://journals.sagepub.com/doi/10.1177/2331216518800869.

Wood, Simon N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." In: *Journal of the Royal Statistical Society (B)* 73.1, pp. 3–36.

Woumans, Evy, Clara D. Martin, Charlotte Vanden Bulcke, Eva Van Assche, Albert Costa, Robert J. Hartsuiker, and Wouter Duyck (2015). "Can faces prime a language?" In: *Psychological science* 26.9, pp. 1343–1352.

Wu, Youyou, David Stillwell, H. Andrew Schwartz, and Michal Kosinski (2017). "Birds of a Feather Do Flock Together: Behavior-Based Personality-Assessment Method Reveals Personality Similarity Among Couples and Friends." In: *Psychological science* 28.3, pp. 276–284.

Zadra, Jonathan R. and Gerald L. Clore (2011). "Emotion and perception: The role of affective information." In: *Wiley interdisciplinary reviews: cognitive science* 2.6, pp. 676–685.

Zekveld, Adriana A., Thomas Koelewijn, and Sophia E. Kramer (2018). "The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge." In: *Trends in Hearing* 22, pp. 1–25. ISSN: 2331-2165. DOI: `10.1177/2331216518777174`. URL: `http://journals.sagepub.com/doi/10.1177/2331216518777174`.

Zeman, Adam, Michaela Dewar, and Sergio Della Sala (2015). "Lives without imagery – Congenital aphantasia." In: *Cortex* 73.June, pp. 129–131. DOI: `10.1016/j.cortex.2015.05.019`. URL: `http://linkinghub.elsevier.com/retrieve/pii/S0010945215001781`.

Zwaan, Rolf A., Carol J. Madden, Richard H. Yaxley, and Mark E. Aveyard (2004). "Moving words: Dynamic representations in language comprehension." In: *Cognitive Science* 28.4, pp. 611–619. ISSN: 03640213. DOI: `10.1016/j.cogsci.2004.03.004`.

Appendix A

Experiment Stimuli

A.1 Morpho-Syntactic Errors

He frequently walk his dog in the park downtown.

She seldom swim three miles at the pool nearby.

You always speaks the truth when I ask you to.

We often sings old songs together at camp.

They normally eats their lunch inside the building.

I seldom buys things for myself or my friends.

He always ride his bike to work in June.

They often buys their milk at the store down the street.

She usually drive her car slowly in the snow.

They rarely wraps any gifts for a friend’s birthday.

She always give her son money for candy.

They often helps their friends with their homework.

I often reads a book on long-haul flights.

He often speak Turkish during recess.

We never cooks dinner at home on Fridays.

We seldom sends postcards from our vacation.

He usually eat bacon in the morning.

She never take the bus to the library.

He constantly drive his car around the city.

She normally fly to Europe around Christmas.

He frequently have burgers for dinner after work.

She always watch shows on her friend's Netflix.

I constantly wears my watch, even at night.

You rarely comes over to my house anymore.

We frequently travels places in the summer.

She sometimes read books on arts and crafts.

He sometimes listen to songs from his childhood.

She usually ride transit to get to school.

A.2 Semantic Anomalies

Bees often collect storage in our backyard.

Cats frequently hunt bricks around their homes.

Students seldom forget dancers in the locker room.

Dancers usually wear limbs at a performance.

Grandma always cleans the cheek by herself.

We always hear skies in the forest.

He never catches firms thrown with a spin.

She rarely attends use in the morning.

People often read heads for pleasure at night.

She sometimes waters terms around the block.

Bikers constantly fear scales in their tires.

Dogs sometimes chase teas on the road for fun.

Bosses normally pay their cups weekly or monthly.

Her mum frequently writes cultures to friends in Europe.

Plants usually need money to grow and bloom.

Lisa rarely cooks chart with rice for dinner.

A.3 Socio-Cultural Clashes

Produced by male speaker:

I usually wear lip gloss to work and at home.

I always enjoy knitting in my free time.

I normally wear high heels to formal parties.

I constantly help my mom with cooking and chores.

I usually fix the holes in my clothes myself.

I sometimes buy my bras at Hudson's Bay.

I normally shave my legs every three days.

I usually avoid tampons as I prefer pads.

I frequently apply perfume in the mornings.

I often drink wine on the weekends.

I often prepare muffins for block parties.

I always enjoy coloring when I am stressed.

I sometimes fight the good fight for the homeless.

I always watch movies on fashion with friends.

I frequently buy new soap for the kitchen.

I constantly wear dresses to board meetings.

I constantly carry a purse with all my things.

I normally wipe the table after we had lunch.

I frequently play Wii games with my friends.

I sometimes grow my hair quite long for fun.

I usually prefer reading over coloring.

I often visit my friends when I feel lonely.

I usually go to the spa to relax after work.

I often wear flowers in my hair for work.

I always wear hair bands to hold in my bangs.

I frequently read gossip news during lunch break.

I often work with kids in my daytime job.

I normally clean the floor with a soft sponge.

I constantly think of home when I travel.

I sometimes shave my arms in the summer months.

Produced by female speaker:

I usually wear blue jeans to work and at home.

I always enjoy football in my free time.

I normally wear dress shoes to formal parties.

I constantly help my dad with garden work.

I usually fix the brakes on my Dodge myself.

I sometimes buy my ties at Hudson's Bay.

I normally shave my beard every three days.

I usually avoid urinals as I prefer stalls.

I frequently apply cologne in the mornings.

I often drink beer on the weekends.

I often prepare spare ribs for block parties.

I always enjoy gaming when I am stressed.

I sometimes fight my brothers over stupid stuff.

I always watch movies about war with my friends.

I frequently buy new coal for the barbecue.

I constantly wear suits to board meetings.

I constantly carry a gun to defend myself.

I normally wipe my hard drive every few months.

I frequently play war games with my buddies.

I sometimes grow my beard quite long for fun.

I usually prefer hockey over football.

I often visit strip clubs when I feel lonely.

I usually go to the bar to relax after work.

I often wear cufflinks on my shirts for work.

I always wear a hat to cover my head.

I frequently read sports news during lunch break.

I often work on cars in my daytime job.

I normally clean the car with a soft sponge.

I constantly think of sex when I am home.

I sometimes shave my chest in the summer months.

A.4 Non-Anomalous Items

A.4.1 Not Dependent on Speaker Gender

He frequently walks his dog in the park downtown.

She seldom swims three miles at the pool nearby.

They often buy their milk at the store down the street.

You always speak the truth when I ask you to.

We often sing old songs together at camp.

They normally eat their lunch inside the building.

I seldom buy things for myself or my friends.

He always rides his bike to work in June.

She usually drives her car slowly in the snow.

They rarely wrap any gifts for a friend's birthday.

She always gives her son money for candy.

They often help their friends with their homework.

I often read a book on long-haul flights.

He often speaks Turkish during recess.

We never cook dinner at home on Fridays.

We seldom send postcards from our vacation.

He usually eats bacon in the morning.

She never takes the bus to the library.

He constantly drives his car around the city.

She normally flies to Europe around Christmas.

He frequently has burgers for dinner after work.

She always watches shows on her friend's Netflix.

I constantly wear my watch, even at night.

You rarely come over to my house anymore.

We frequently travel places in the summer.

She sometimes reads books on arts and crafts.

He sometimes listens to songs from his childhood.

She usually rides transit to get to school.

Cats frequently hunt mice around their homes.

Students seldom forget towels in the locker room.

Dancers usually wear skirts at a performance.

Grandma always cleans the house by herself.

We always hear birds in the forest.

He never catches balls thrown with a spin.

She rarely attends class in the morning.

People often read books for pleasure at night.

She sometimes waters plants around the block.

Bikers constantly fear flats on their tires.

Dogs sometimes chase bikes on the road for fun.

Bosses normally pay their staff weekly or monthly.

Her mum frequently writes letters to friends in Europe.

Plants usually need water to grow and bloom.

Lisa rarely cooks fish with rice for dinner.

Cats often catch mice in the night.

Dogs generally fetch sticks on a walk.

Rabbits often change colour in the winter.

Seagulls often catch fish in groups.

Horses generally accept being trained.

Birds often sing loudly in the morning.

Foxes usually live alone in the forest.

Wolves seldom enjoy being near humans.

Chickens normally live in a coop.

Roosters often announce the sunrise.

Bears sometimes approach picknickers.

Moose sometimes block the roads in a park.

Geese always watch out for their young.

Goslings generally follow their mother.

Ants usually build large anthills.

Bees generally make honey from pollen.

Bugs often live where it is humid.

Skunks often spray their enemies.

Donkeys often carry heavy loads.

Rabbits generally enjoy eating carrots.

Dogs typically bark at intruders.

Cats often purr when they are content.

Birds frequently sit perched atop trees.

Chickens generally lay many eggs.

Hippos never jump high into the air.

Giraffes always have very long necks.

Rhinos usually have large horns.

Beavers often build dams in rivers.

Monkeys generally enjoy bananas.

Lions frequently roar very loudly.

Leopards generally run very fast.

Bears seldom venture into cities.

A.4.2 Dependent on Speaker Gender

Produced by male speaker:

I usually wear blue jeans to work and at home.

I always enjoy football in my free time.

I normally wear dress shoes to formal parties.

I constantly help my dad with garden work.

I usually fix the brakes on my Dodge myself.

I sometimes buy my ties at Hudson's Bay.

I normally shave my beard every three days.

I usually avoid urinals as I prefer stalls.

I frequently apply cologne in the mornings.

I often drink beer on the weekends.

I often prepare spare ribs for block parties.

I always enjoy gaming when I am stressed.

I sometimes fight my brothers over stupid stuff.

I always watch movies about war with my friends.

I frequently buy new coal for the barbecue.

I constantly wear suits to board meetings.

I constantly carry a gun to defend myself.

I normally wipe my hard drive every few months.

I frequently play war games with my buddies.

I sometimes grow my beard quite long for fun.

I usually prefer hockey over football.

I often visit strip clubs when I feel lonely.

I usually go to the bar to relax after work.

I often wear cufflinks on my shirts for work.

I always wear a hat to cover my head.

I frequently read sports news during lunch break.

I often work on cars in my daytime job.

I normally clean the car with a soft sponge.

I constantly think of sex when I am home.

I sometimes shave my chest in the summer months.

Produced by female speaker:

I usually wear lip gloss to work and at home.

I always enjoy knitting in my free time.

I normally wear high heels to formal parties.

I constantly help my mom with cooking and chores.

I usually fix the holes in my clothes myself.

I sometimes buy my bras at Hudson's Bay.

I normally shave my legs every three days.

I usually avoid tampons as I prefer pads.

I frequently apply perfume in the mornings.

I often drink wine on the weekends.

I often prepare muffins for block parties.

I always enjoy coloring when I am stressed.

I sometimes fight the good fight for the homeless.

I always watch movies on fashion with friends.

I frequently buy new soap for the kitchen.

I constantly wear dresses to board meetings.

I constantly carry a purse with all my things.

I normally wipe the table after we had lunch.

I frequently play Wii games with my friends.

I sometimes grow my hair quite long for fun.

I usually prefer reading over coloring.

I often visit my friends when I feel lonely.

I usually go to the spa to relax after work.

I often wear flowers in my hair for work.

I always wear hair bands to hold in my bangs.

I frequently read gossip news during lunch break.

I often work with kids in my daytime job.

I normally clean the floor with a soft sponge.

I constantly think of home when I travel.

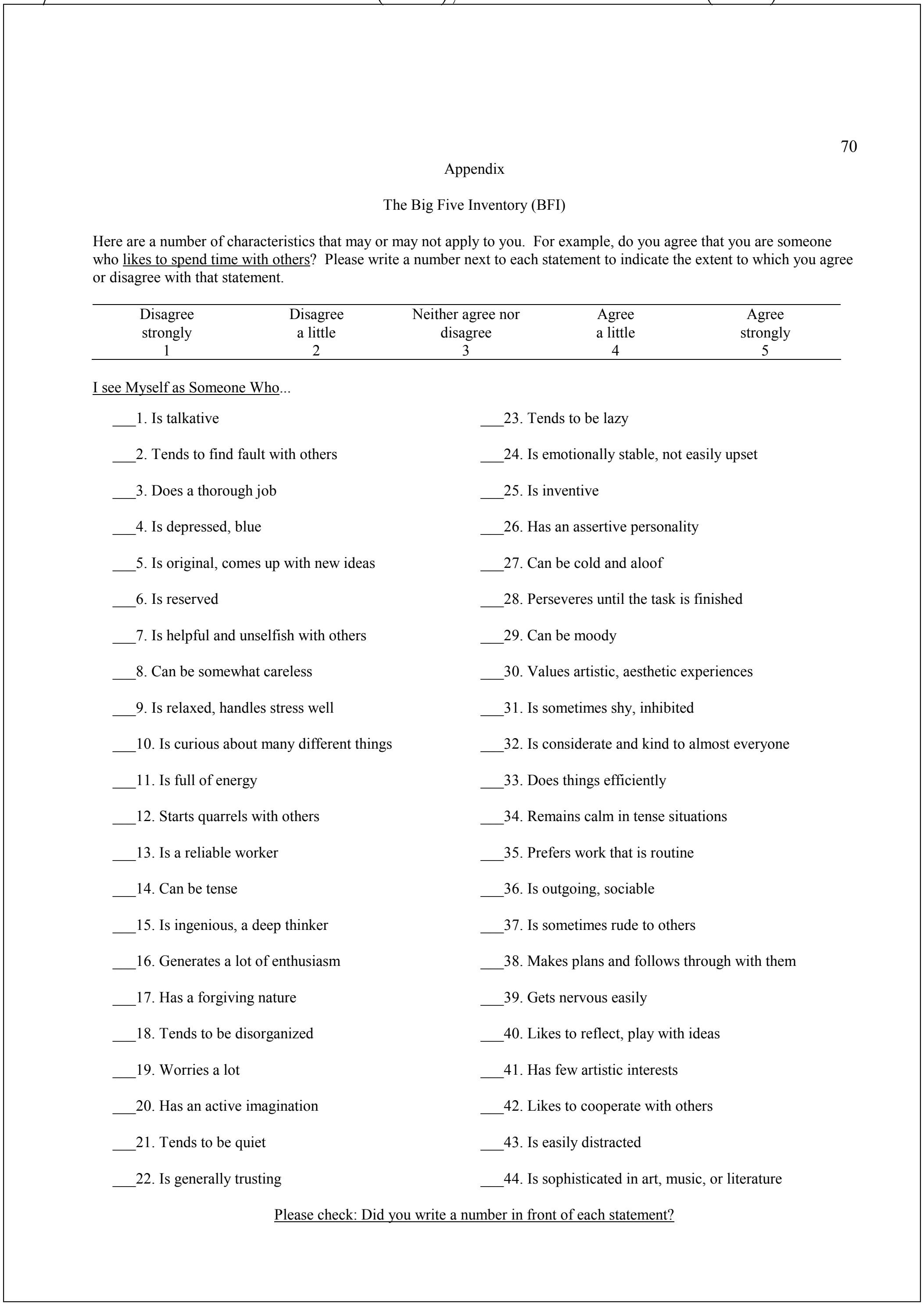
I sometimes shave my arms in the summer months.

Appendix B

Post-Tests

B.1 Big Five Personality Assessment

from p. 70/71 in John and Srivastava (1999); see also John et al. (1991).



Appendix

The Big Five Inventory (BFI)

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
1	2	3	4	5

I see Myself as Someone Who...

___1. Is talkative

___2. Tends to find fault with others

___3. Does a thorough job

___4. Is depressed, blue

___5. Is original, comes up with new ideas

___6. Is reserved

___7. Is helpful and unselfish with others

___8. Can be somewhat careless

___9. Is relaxed, handles stress well

___10. Is curious about many different things

___11. Is full of energy

___12. Starts quarrels with others

___13. Is a reliable worker

___14. Can be tense

___15. Is ingenious, a deep thinker

___16. Generates a lot of enthusiasm

___17. Has a forgiving nature

___18. Tends to be disorganized

___19. Worries a lot

___20. Has an active imagination

___21. Tends to be quiet

___22. Is generally trusting

___23. Tends to be lazy

___24. Is emotionally stable, not easily upset

___25. Is inventive

___26. Has an assertive personality

___27. Can be cold and aloof

___28. Perseveres until the task is finished

___29. Can be moody

___30. Values artistic, aesthetic experiences

___31. Is sometimes shy, inhibited

___32. Is considerate and kind to almost everyone

___33. Does things efficiently

___34. Remains calm in tense situations

___35. Prefers work that is routine

___36. Is outgoing, sociable

___37. Is sometimes rude to others

___38. Makes plans and follows through with them

___39. Gets nervous easily

___40. Likes to reflect, play with ideas

___41. Has few artistic interests

___42. Likes to cooperate with others

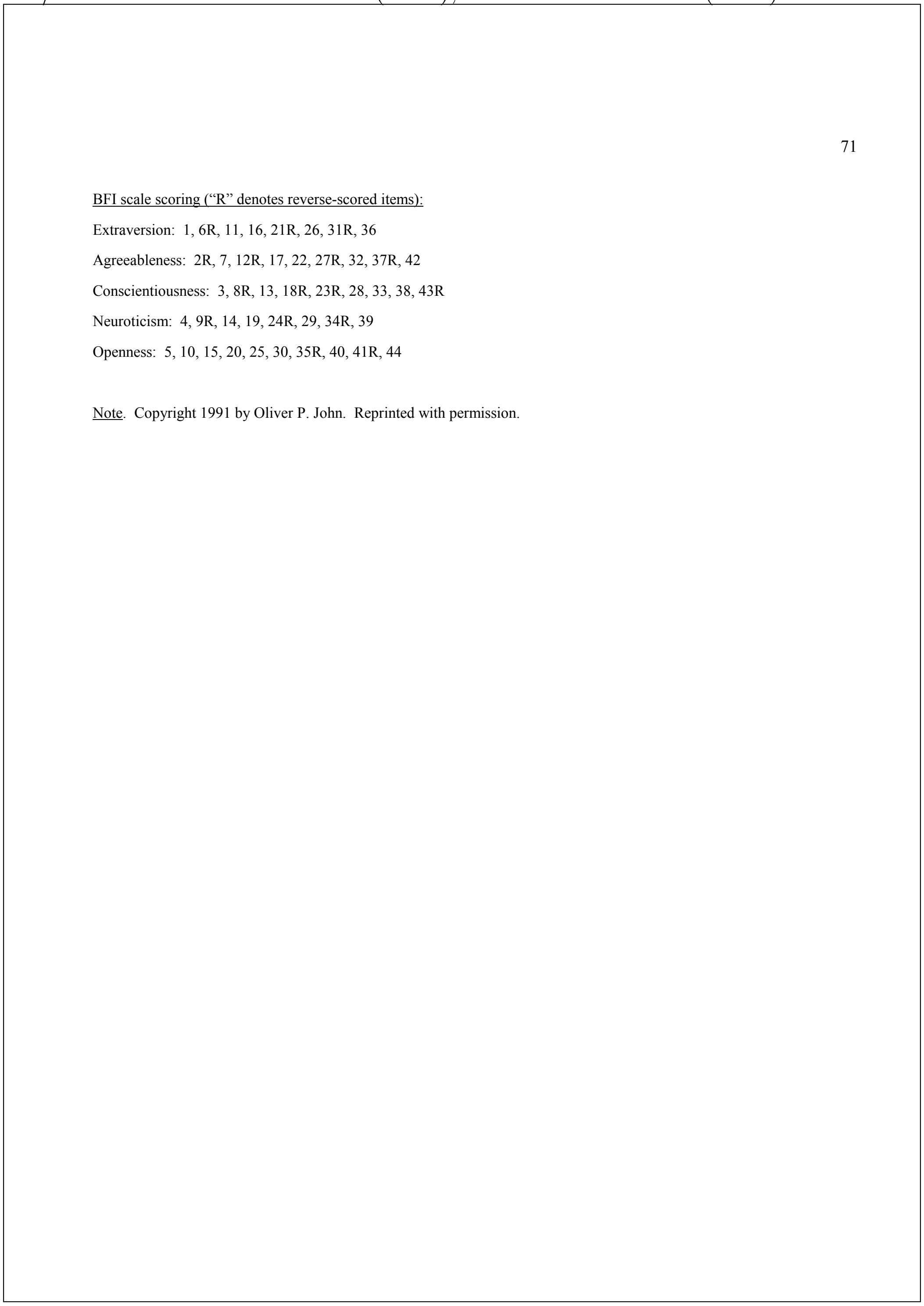
___43. Is easily distracted

___44. Is sophisticated in art, music, or literature

Please check: Did you write a number in front of each statement?

B.2 Big Five Personality Assessment

from p. 70/71 in John and Srivastava (1999); see also John et al. (1991).



71

BFI scale scoring ("R" denotes reverse-scored items):

Extraversion: 1, 6R, 11, 16, 21R, 26, 31R, 36

Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42

Conscientiousness: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R

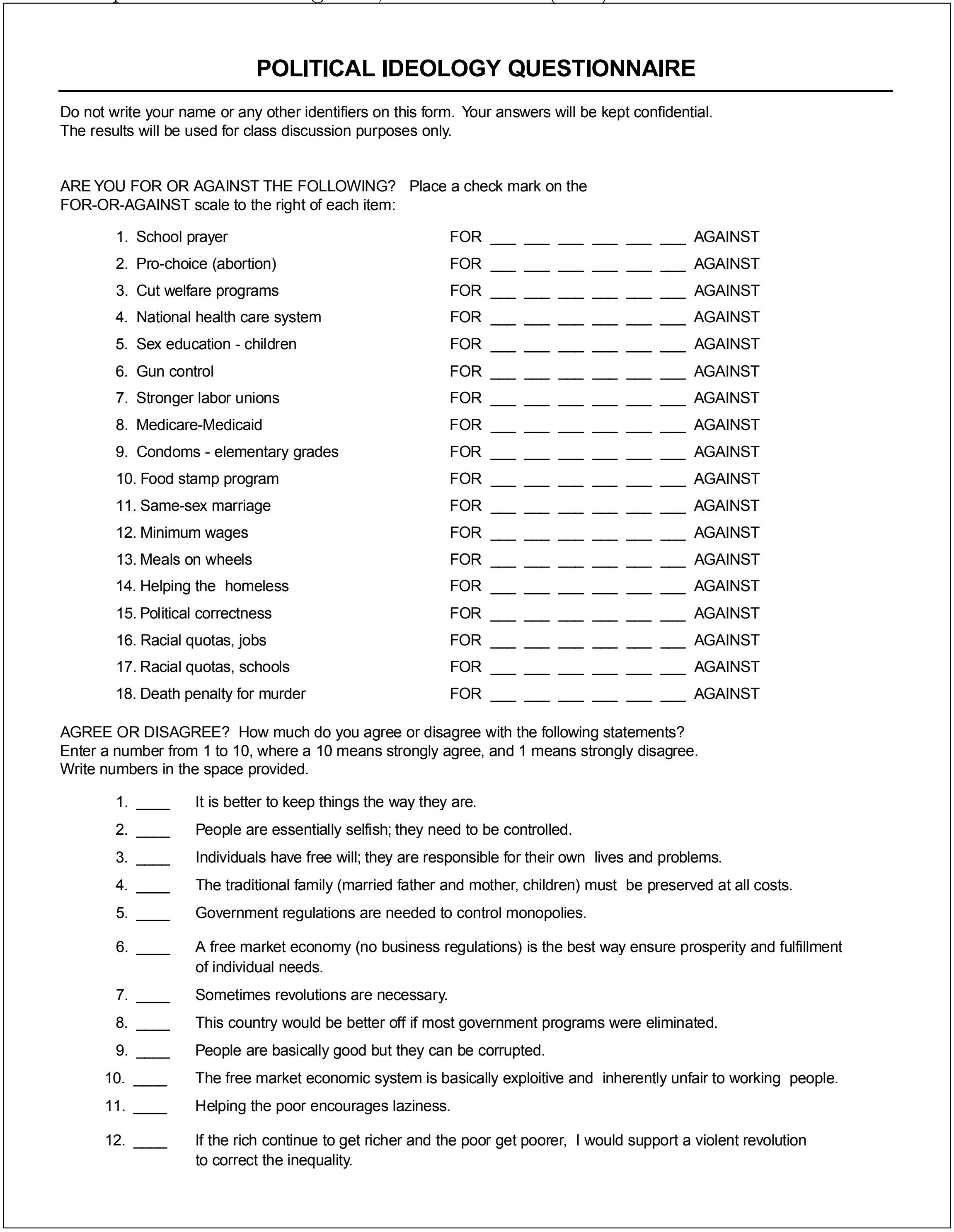
Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39

Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44

Note. Copyright 1991 by Oliver P. John. Reprinted with permission.

B.3 Political Values Questionnaire

as used in Experiments I through III; from Grenier (n.d.).



POLITICAL IDEOLOGY QUESTIONNAIRE

Do not write your name or any other identifiers on this form. Your answers will be kept confidential.
The results will be used for class discussion purposes only.

ARE YOU FOR OR AGAINST THE FOLLOWING? Place a check mark on the FOR-OR-AGAINST scale to the right of each item:

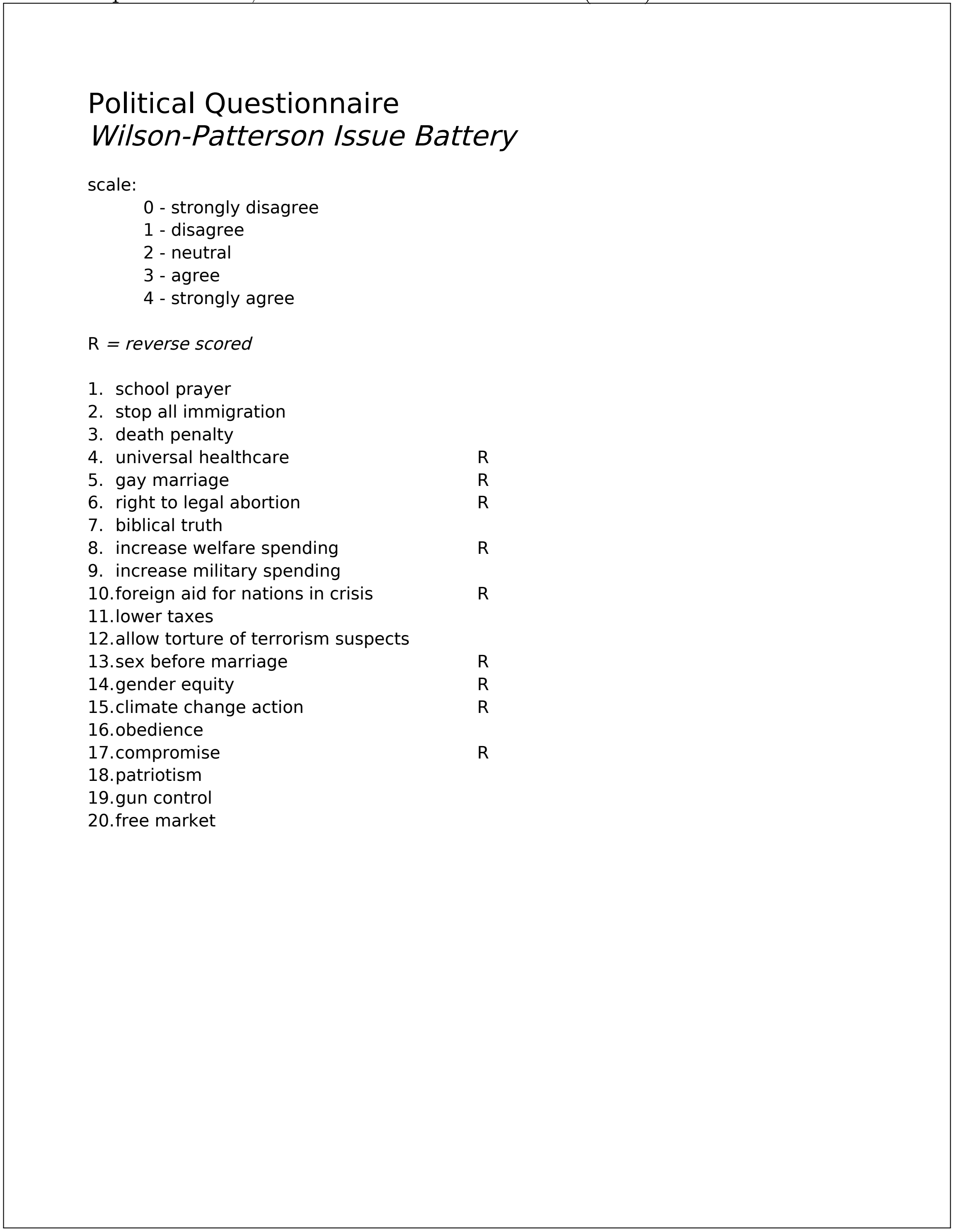
1. School prayer FOR ___ ___ ___ ___ ___ ___ AGAINST
2. Pro-choice (abortion) FOR ___ ___ ___ ___ ___ ___ AGAINST
3. Cut welfare programs FOR ___ ___ ___ ___ ___ ___ AGAINST
4. National health care system FOR ___ ___ ___ ___ ___ ___ AGAINST
5. Sex education - children FOR ___ ___ ___ ___ ___ ___ AGAINST
6. Gun control FOR ___ ___ ___ ___ ___ ___ AGAINST
7. Stronger labor unions FOR ___ ___ ___ ___ ___ ___ AGAINST
8. Medicare-Medicaid FOR ___ ___ ___ ___ ___ ___ AGAINST
9. Condoms - elementary grades FOR ___ ___ ___ ___ ___ ___ AGAINST
10. Food stamp program FOR ___ ___ ___ ___ ___ ___ AGAINST
11. Same-sex marriage FOR ___ ___ ___ ___ ___ ___ AGAINST
12. Minimum wages FOR ___ ___ ___ ___ ___ ___ AGAINST
13. Meals on wheels FOR ___ ___ ___ ___ ___ ___ AGAINST
14. Helping the homeless FOR ___ ___ ___ ___ ___ ___ AGAINST
15. Political correctness FOR ___ ___ ___ ___ ___ ___ AGAINST
16. Racial quotas, jobs FOR ___ ___ ___ ___ ___ ___ AGAINST
17. Racial quotas, schools FOR ___ ___ ___ ___ ___ ___ AGAINST
18. Death penalty for murder FOR ___ ___ ___ ___ ___ ___ AGAINST

AGREE OR DISAGREE? How much do you agree or disagree with the following statements?
Enter a number from 1 to 10, where a 10 means strongly agree, and 1 means strongly disagree.
Write numbers in the space provided.

1. ____ It is better to keep things the way they are.
2. ____ People are essentially selfish; they need to be controlled.
3. ____ Individuals have free will; they are responsible for their own lives and problems.
4. ____ The traditional family (married father and mother, children) must be preserved at all costs.
5. ____ Government regulations are needed to control monopolies.
6. ____ A free market economy (no business regulations) is the best way ensure prosperity and fulfillment of individual needs.
7. ____ Sometimes revolutions are necessary.
8. ____ This country would be better off if most government programs were eliminated.
9. ____ People are basically good but they can be corrupted.
10. ____ The free market economic system is basically exploitive and inherently unfair to working people.
11. ____ Helping the poor encourages laziness.
12. ____ If the rich continue to get richer and the poor get poorer, I would support a violent revolution to correct the inequality.

B.4 Wilson-Patterson Political Values Questionnaire

as used in Experiment IV; from Wilson and Patterson (1968).



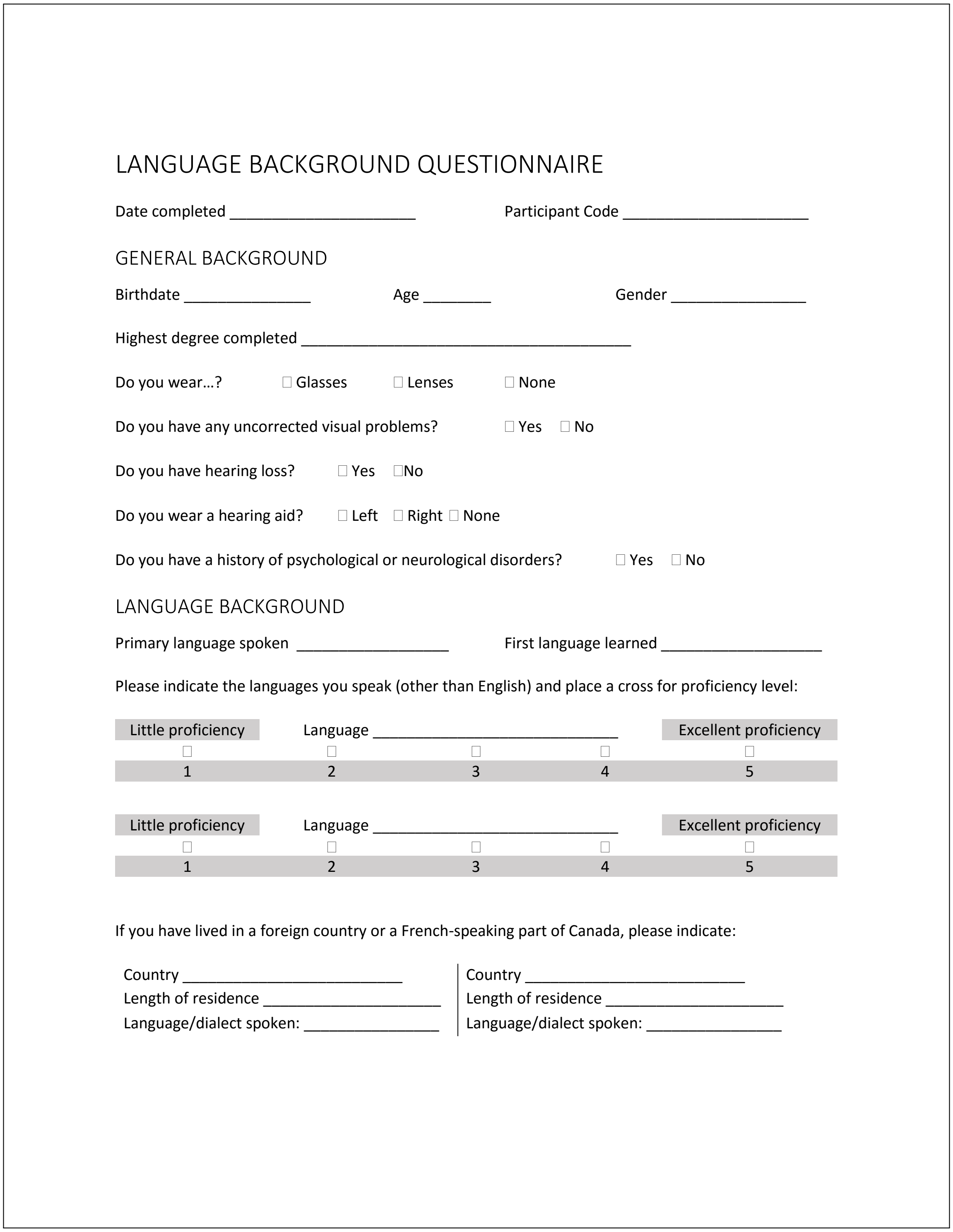
Political Questionnaire
Wilson-Patterson Issue Battery

scale:
0 - strongly disagree
1 - disagree
2 - neutral
3 - agree
4 - strongly agree

R *= reverse scored*

1. school prayer	
2. stop all immigration	
3. death penalty	
4. universal healthcare	R
5. gay marriage	R
6. right to legal abortion	R
7. biblical truth	
8. increase welfare spending	R
9. increase military spending	
10.foreign aid for nations in crisis	R
11.lower taxes	
12.allow torture of terrorism suspects	
13.sex before marriage	R
14.gender equity	R
15.climate change action	R
16.obedience	
17.compromise	R
18.patriotism	
19.gun control	
20.free market	

B.5 Language Background Questionnaire



LANGUAGE BACKGROUND QUESTIONNAIRE

Date completed ______________________ Participant Code ______________________

GENERAL BACKGROUND

Birthdate _______________ Age ________ Gender ________________

Highest degree completed _______________________________________

Do you wear…? □ Glasses □ Lenses □ None

Do you have any uncorrected visual problems? □ Yes □ No

Do you have hearing loss? □ Yes □No

Do you wear a hearing aid? □ Left □ Right □ None

Do you have a history of psychological or neurological disorders? □ Yes □ No

LANGUAGE BACKGROUND

Primary language spoken __________________ First language learned ___________________

Please indicate the languages you speak (other than English) and place a cross for proficiency level:

Little proficiency	Language _____________________________			Excellent proficiency
□	□	□	□	□
1	2	3	4	5

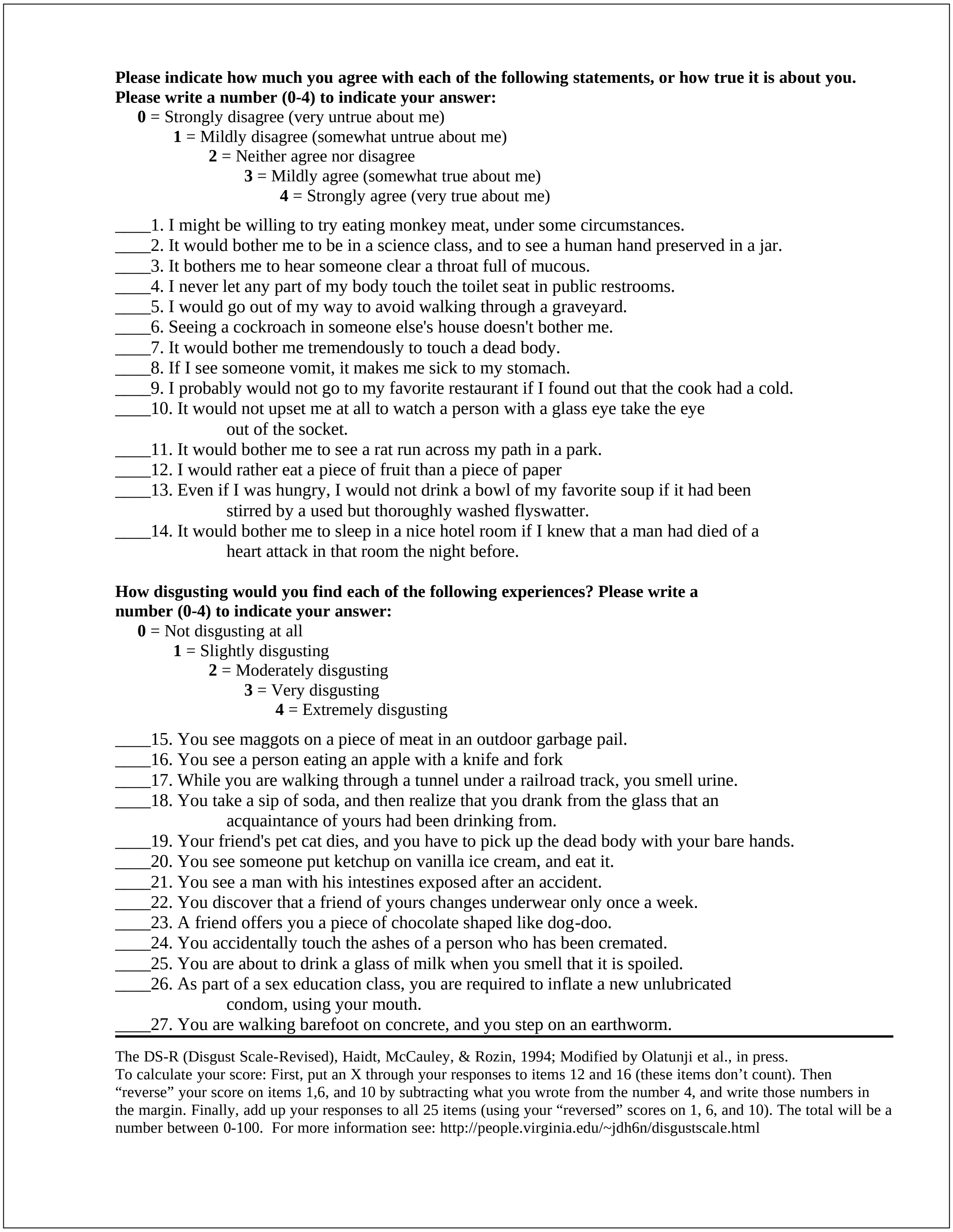
Little proficiency	Language _____________________________			Excellent proficiency
□	□	□	□	□
1	2	3	4	5

If you have lived in a foreign country or a French-speaking part of Canada, please indicate:

Country __________________________
Length of residence _____________________
Language/dialect spoken: ________________

Country __________________________
Length of residence _____________________
Language/dialect spoken: ________________

B.6 Disgust Scale - Revised



Please indicate how much you agree with each of the following statements, or how true it is about you. Please write a number (0-4) to indicate your answer:

0 = Strongly disagree (very untrue about me)
1 = Mildly disagree (somewhat untrue about me)
2 = Neither agree nor disagree
3 = Mildly agree (somewhat true about me)
4 = Strongly agree (very true about me)

____1. I might be willing to try eating monkey meat, under some circumstances.
____2. It would bother me to be in a science class, and to see a human hand preserved in a jar.
____3. It bothers me to hear someone clear a throat full of mucous.
____4. I never let any part of my body touch the toilet seat in public restrooms.
____5. I would go out of my way to avoid walking through a graveyard.
____6. Seeing a cockroach in someone else's house doesn't bother me.
____7. It would bother me tremendously to touch a dead body.
____8. If I see someone vomit, it makes me sick to my stomach.
____9. I probably would not go to my favorite restaurant if I found out that the cook had a cold.
____10. It would not upset me at all to watch a person with a glass eye take the eye out of the socket.
____11. It would bother me to see a rat run across my path in a park.
____12. I would rather eat a piece of fruit than a piece of paper
____13. Even if I was hungry, I would not drink a bowl of my favorite soup if it had been stirred by a used but thoroughly washed flyswatter.
____14. It would bother me to sleep in a nice hotel room if I knew that a man had died of a heart attack in that room the night before.

How disgusting would you find each of the following experiences? Please write a number (0-4) to indicate your answer:

0 = Not disgusting at all
1 = Slightly disgusting
2 = Moderately disgusting
3 = Very disgusting
4 = Extremely disgusting

____15. You see maggots on a piece of meat in an outdoor garbage pail.
____16. You see a person eating an apple with a knife and fork
____17. While you are walking through a tunnel under a railroad track, you smell urine.
____18. You take a sip of soda, and then realize that you drank from the glass that an acquaintance of yours had been drinking from.
____19. Your friend's pet cat dies, and you have to pick up the dead body with your bare hands.
____20. You see someone put ketchup on vanilla ice cream, and eat it.
____21. You see a man with his intestines exposed after an accident.
____22. You discover that a friend of yours changes underwear only once a week.
____23. A friend offers you a piece of chocolate shaped like dog-doo.
____24. You accidentally touch the ashes of a person who has been cremated.
____25. You are about to drink a glass of milk when you smell that it is spoiled.
____26. As part of a sex education class, you are required to inflate a new unlubricated condom, using your mouth.
____27. You are walking barefoot on concrete, and you step on an earthworm.

The DS-R (Disgust Scale-Revised), Haidt, McCauley, & Rozin, 1994; Modified by Olatunji et al., in press.
To calculate your score: First, put an X through your responses to items 12 and 16 (these items don't count). Then "reverse" your score on items 1,6, and 10 by subtracting what you wrote from the number 4, and write those numbers in the margin. Finally, add up your responses to all 25 items (using your "reversed" scores on 1, 6, and 10). The total will be a number between 0-100. For more information see: http://people.virginia.edu/~jdh6n/disgustscale.html