# The processing of stereotype-relevant information during reading

**Christine Berta Maria Häcker**

A thesis submitted to the University of Birmingham for the degree of Doctor of Philosophy

**School of Psychology**
**University of Birmingham**
**September 2009**

# Abstract

I examined the processing of stereotype-relevant information during reading, in particular the degree to which stereotype-mismatch detection and resolution are resource-dependent. In addition I investigated the effects of stereotype-relevant episodic representations on subsequent linguistic and non-linguistic processing. Experiment 1 showed that reading participants looked longer at pronouns that mismatched the stereotypical gender of the agent than at stereotype-matching pronouns (e.g., "...the *secretary* familiarised *herself/ himself*..."). Experiment 1 also showed that mismatch detection can take place even when readers are cognitively busy, but that later integration processes might be compromised, resulting in an increased memory bias. Experiments 2 and 3 showed that the episodic representations resulting from reading stereotype-relevant sentences are strong and stable enough to cancel out a mismatch effect in a second sentence, unless the stereotypical representation is reemphasised by a repetition of the occupation label. Experiments 4 to 7 showed that gender-categorisation was facilitated for target faces that matched rather than mismatched a priming stereotypical occupation label (e.g., secretary); such an effect was not found for more complex prime stereotype-relevant sentences. It can be concluded that episodic stereotype-relevant representations can affect further processing of linguistic and non-linguistic information. This influence, however, is limited by existing stereotype representations.

# Acknowledgments

I would like to extend my most heartfelt thanks to my supervisors Prof. Antje Meyer and Dr. Kimberly Quinn for their guidance and support throughout my PhD project and in completing this thesis. I honestly could have not wished for better supervisors. I am most grateful for a PhD studentship awarded to me by the School of Psychology, for making this thesis possible. I am also grateful to Dr. Steven Frisson, Dr. Mike Harris, Dr. Dietmar Heinke and Dr. Andrew Olson for their advice on aspects of methodology and analysis. I would also like to thank all former and present members of the psycholinguistic and social research groups as well as many other colleagues for their most valued support and friendship (you know who you are!). Thanks also to all members of the Birmingham University Buddhist Society for helping me to keep my sanity throughout these years. Last, but not least, I would like to thank my family and friends outside University for their love, inspiration, patience and support.

## Contents

## Appendix

## Tables

## Figures

# Chapter 1
# Introduction

# 1. Motivation

On 4 May 2009, the BBC news reported, “Labour's deputy leader denies a report she would fight for the leadership, amid speculation over Gordon Brown's position”. Readers might have been surprised when they read the word “she” in this sentence, because they might have assumed that Labour’s deputy leader was male on the basis of most leading figures in politics being male. Similarly, people might find it also difficult to solve the following riddle: “A father and son are involved in a horrific car accident. The father is killed, and the son is rushed to hospital for emergency surgery. Upon their arrival, however, the surgeon takes one look at the child and says, ‘I cannot operate on him. He is my son.’ How is this possible?” Of course, the simple answer is that the surgeon is the child’s mother. The reason why people might have problems with finding this solution is that most surgeons are male. These examples illustrate that some occupations are by default assumed to be held by men whereas others are assumed to be held by women. As for many occupations, there are no reasons why the other gender should not be able to fulfill the occupation; these assumptions are stereotypical overgeneralisations. As such, they play an important part in the processing of social information by simplifying and organising it. They influence how people perceive other people, encode information about them, reason about them, and judge them, which information they remember about them and how they behave towards them (Hilton & von Hippel, 1996). For the most part, people are unaware of using stereotypes and can therefore not account for their influences on the impressions they form of other people.

The motivation for the present research was to contribute to the knowledge about the processing of stereotype-relevant information during reading. What happens, for example, when readers encounter stereotype-mismatching information like about the female labour deputy leader? Do they only notice this kind of information when their attention is undivided or also when they are busy doing something else while reading? Do they remember later that the deputy leader was a woman? When they read about this person again within the same context, are they surprised again when they encounter another piece of gender-specifying information? Could reading about a female deputy leader make readers process female or male faces differently?

In order to investigate such questions, I used stereotype-matching sentences such as “Last week the *secretary* familiarised *herself* with the new photocopier” and stereotype-mismatching sentences such as “Last week the *secretary* familiarised *himself* with the new photocopier” (emphases added). It has commonly been found with this kind of sentences that readers have processing difficulties when encountering the mismatching pronoun, reflected in longer self-paced reading times for mismatching than matching sentences (e.g., Carreiras, Garnham, Oakhill, & Cain, 1996) and longer gaze durations on mismatching than matching pronouns (e.g., Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Sturt, 2003). I will refer to this effect from now on as *mismatch effect*.

My thesis comprised seven experiments. Experiment 1 had several goals: (1) to replicate the previously found mismatch effect as a basis for its further investigation; (2) to examine what participants remembered of the stereotype-relevant information in order to gain insight into the representations they constructed during online processing; and (3) to examine the effect of cognitive load on both online processing of and memory for stereotype-relevant information, to determine whether stereotype-mismatching information is only detected and resolved when readers’ cognitive capacities are plentiful or whether these processes also take place when readers are cognitively busy. To tackle these goals, I monitored participants' eye movements to measure online reading times for stereotype-matching and -mismatching sentences. Half of the participants carried out a cognitive load before reading the sentences (a 5-digit number retention task). Following the reading task, I administered a questionnaire asking participants whether the occupations mentioned in the sentences had been held by women or men.

The goal of Experiments 2 and 3 was to investigate whether the episodic representations constructed during reading stereotype-relevant information have an effect on subsequent linguistic processing. I sought to examine whether the episodic representations would be strong and stable enough to override the stereotypical representation and therefore cancel out a second effect of stereotype-mismatch in the further discourse context. I studied whether such a cancelation effect would occur only for the further processing of the same member of a stereotyped group (token) or whether it would generalise to other members of the category (type). Participants in Experiment 2 read sentence pairs. The first sentences were similar to the ones in Experiments 1; the second sentences repeated the role name and pronoun information and referred to either the same or a different agent (token and type conditions, respectively). An example of a sentence pair in the token condition was: "The *elderly secretary* thoroughly familiarised *herself/himself* with the new computer a few months before retiring. To everyone's surprise, the *secretary* really enjoyed *herself/himself* while exploring the potential of the computer". An example sentence pair in the type condition was: "The *elderly secretary* reluctantly familiarised *herself/himself* with the new computer a few months before retiring. In contrast, the *new secretary* really enjoyed *herself/himself* while exploring the potential of the computer". The results showed that a mismatch effect was observed for both the first and second sentences.

The source of the second mismatch effect was unclear. It could have arisen because the episodic representations constructed during reading the first sentence were not strong or stable enough to override the gender-stereotypical representations. However, it could also have arisen because the repetition of the occupation label in the second sentence gave rise to repeated stereotype activation. To disambiguate the source of the mismatch effect in the second sentence, the token condition was tested again in Experiment 3. Here, instead of explicitly referring back to the agent, reference was left implicit (e.g., "The elderly *secretary* thoroughly familiarised *herself/himself* with the new computer a few months before retiring and, to everyone's surprise, really enjoyed *herself/himself* while exploring the potential of the computer").

The goal of Experiments 4 to 7 was to examine whether the episodic representations constructed during reading stereotype-relevant information have an influence beyond linguistic processing. I therefore combined the reading task with a non-linguistic probe task featuring pictures of female and male faces. I sought to determine whether gender-categorisation latencies for the faces would be facilitated by matching compared to mismatching stereotype-relevant linguistic gender information. In Experiment 4 to 6, I also investigated whether the reflexive pronouns, which were always the most recent gender-relevant information in the sentences, had an additional facilitation effect on the picture categorisation times. I therefore compared the picture-categorisation times following sentences with matching agent-pronoun combinations (e.g., "Last week the *secretary* familiarised *herself* with the new photocopier") to the picture-categorisation times following sentences that were similar in meaning but did not include a reflexive pronoun ("Last week the *secretary* became familiarised with the new photocopier"). In Experiment 7, I used as stimuli in the reading task bare gender-stereotypical nouns (e.g., "secretary", "mechanic").

The mismatch effect has been investigated in a number of psycholinguistics studies (Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Sturt, 2003), but this past research has typically focused on the mere activation of stereotypes during reading. The novel contributions of the present research are to examine the resource-dependency of this process and the representations that result from reading stereotype-relevant information as an interpretative framework for continued processing and memory of social information; throughout the thesis, I frame these contributions in terms of a working model which includes episodic exemplar representations of specific processing events and semantic prototype representations of pre-experimentally existing stereotypical knowledge. From a psycholinguistics perspective this research contributes to an understanding of the local, online-processing of and memory for stereotypical gender violations as well as their integration into wider discourse representations. From a broader social perspective, this research might contribute to an understanding of the stability and scope of stereotype-driven processing.

Next, the psycholinguistic background to this research will be outlined, followed by an overview of my working model, and finally an overview of the thesis.

# 2. Psycholinguistic background

## 2.1 The reading process

Participants in my experiments read words and sentences. I will therefore give a brief overview of selected theoretical approaches to the processes involved in reading. These processes comprise recognising single letters, recognising words and accessing their meaning, analysing the syntactic sentence structure (parsing) and, finally, deriving the sentence meaning (Harley, 2001).

Many studies and theories on single-letter and word recognition (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Finkbeiner & Coltheart, 2009; Grainger & Jacobs, 1996; Tydgat & Grainger, 2009) are based on the seminal work of McClelland and Rumelhart (1981). According to their interactive activation model, word recognition covers the levels of visual feature detection, letter detection and word detection. These levels are connected through excitatory and inhibitory pathways. When readers process a written word, for instance "TIME", the feature detectors for the first letter position activate a vertical and a horizontal line, while inhibiting other features. On the letter level, this feature activation excites the letter pattern "T", while inhibiting other letters. On the word level, the letter activation excites all four-letter words, starting with the letter "T", while inhibiting words starting with other letters as well as shorter or longer words beginning with the letter "T". These bottom-up processes are accompanied by top-down processes that help, for example, in selecting the remaining letters in the word by limiting the number of reasonable choices, which in return facilitates the activation of certain features.

Once the orthographical form of a word has been identified, its meaning can be accessed. Access to meaning is widely assumed to lead over the phonological form (e.g., Ashby & Martin, 2008; Frost, 1998; Perfetti, Bell, & Delaney, 1988; Ziegler & Jacobs, 1995; but see Baron 1973; Bower, 1970; Piras & Marangolo, 2004).

The process of readers arriving at the phonetic form is often modelled via a dual-route-access (e.g., Baron & Strawson, 1976; Coltheart, 1978; Coltheart, Curtis, Atkins, & Haller, 1993; Marshall & Newcombe, 1973; but see Share, 2008). One route is assumed to lead over the rule-based mapping of letters or letter clusters to the corresponding sounds. However, many English words are irregular and cannot be assembled in this way. Therefore, a second route is assumed for irregular words. This direct route leads straight from the visual input to the phonological representation stored in memory. As an alternative, Seidenberg and McClelland (1989) have proposed a single-route connectionist model with orthographic input units and phonological output units, connected over hidden units. The weights of the connections between the units are adapted during the training of the model which over time leads to the appropriate input-output pairings (see also Plaut & McClelland, 1993; Plaut, McClelland, Seidenberg, & Patterson, 1996; Yap & Balota, 2009). Some authors state that in order to retrieve word meaning from the mental lexicon, the phonological form has to be accessed (e.g., Frost, 1998; Lukatela & Turvey, 1994; van Orden, Pennington, & Stone, 1990); others, however, claim that the orthographic representation is sufficient (e.g., Coltheart, Rastle, Perry, Langdon, & Zeigler, 2001; Zorzi, Houghton, & Butterworth, 1998).

There are a number of theories about how the mental lexicon is organised and how it is accessed. Some theorists believe that the units of representation in the lexicon are morphemes like {in-} + {considerate} + {-ly} (e.g., Taft & Ardasinski, 2006; Taft & Forster, 1975), whereas others believe that they are words like {inconsiderately} (e.g., Fowler, Napps, & Feldman, 1985; Giraudo & Grainger, 2000; Kempley & Morton, 1982; Lukatela, Gligorijevic, Kostic, & Turvey, 1980). Some attempts have been made to integrate both approaches into parallel dual-route models (e.g., Baayen, Dijkstra, & Schreuder, 1997).

A distinction more relevant to this thesis is that between the representation of linguistic aspects and extralinguistic aspects of meaning (e.g., Bierwisch & Schreuder, 1992; Levelt, Roelofs, & Meyer, 1999; Wiese, 1999; Viggliocco & Vinson, 2007; Wiese, 2004). For example, Wiese (1999) distinguishes between the semantic system and the conceptual system. The semantic system is viewed as “accounting for those aspects of meaning that have reflexes in the linguistic system and is part of language" (p. 200), whereas the conceptual system is viewed as capturing aspects of meaning that "do not enter lexical information directly, but only in the form of their semantic 'proxies'" (p. 199). In Wiese's model, the conceptual and semantic systems form one module (semantic-conceptual module), which is connected to the phonetic-phonological module via the syntactic module. Levelt and colleagues (1999)[1] also distinguish between a conceptual level, containing (lexical) concepts, and a level of lemmas, which encode syntactic information about the words (Levelt, 2001). The conceptual and lemma level are connected to the phonological level via spreading activation.

Looking more closely at the conceptual level, there are many theories about its organisation. Very broadly, these theories can be divided into holistic and decompositional representations (Caramazza, 1997; Smith, 1998; Vigliocco & Vinson, 2007). Within the holistic approach, each concept (e.g., dog) is represented as a single unit, which is connected with other concept units (e.g., fur, cat). Within the decompositional approach, each concept is represented by a number of units or features that individually do not have a meaningful interpretation (see Conrey & Smith, 2007). The organisation of word meanings as holistic units has, for example, been modelled as an associative network, where word meaning units are connected over associative links (e.g., Anderson & Bower, 1973; Carlston, 1994; Collins & Loftus, 1975; Collins & Quillian, 1969). Another holistic theory views conceptual

[1] Although the theory of lexical access was designed as a speech production model, it can be extended to reading either by postulating modality-specific lemmas or amodal lemmas with links to modality-specific lexemes (see Roelofs, Meyer, & Levelt, 1998, pp. 222, 228).

representations as prototypes (e.g., Homa, 1984; Reed, 1972; Rosch & Mervis, 1975; see also Minda & Smith, 2001). Prototypes are abstract, representative members of a category and are often assumed to contain all features representing the generic knowledge about an object, concept, person or group (Smith, 1998). In contrast to this, exemplar models view representations as particular stored instances that represent specific information about a particular stimulus or instance (e.g., Brooks, 1978; Hintzman, 1986; Jacoby & Brooks, 1984; Medin & Schaffer, 1978; Nosofsky, 1986; see also Nosofsky & Zaki, 2002).

An early decompositional idea about word meaning representations is one of semantic features (e.g., McNamara & Miller 1989; Schank, 1972; Wilks, 1976; see also Vinson & Viggliocco, 2008), where word meaning is represented as a combination of such features. “Woman”, for example, could then be represented as [+HUMAN], [+FEMALE], [+ADULT], whereas “girl” could be represented as [+HUMAN], [+FEMALE], [-ADULT]. Many current decompositional theories are modelled in a connectionist way (e.g., Chang, Dell, & Bock, 2006; Devlin, Gonnerman, Andersen, & Seidenberg, 1998; Farah & McClelland, 1991; Hinton & Shallice, 1991; McRae, deSa, & Seidenberg, 1997; Plaut, 1995; Vigliocco, Vinson, Lewis, & Garrett, 2004). Within connectionist models, concepts are represented in a distributed way as activation patterns within a network of interconnected units. The units are linked by weighted connections that determine how much activation spreads through them. Different patterns of activation throughout the network represent different concepts (for overviews, see Clark, 1993; Harley, 2001; Smith, 1998).

For successful reading, it is necessary not only to access the individual word meanings, but also to combine words into a grammatical structure. Mitchell (1994) divided this process of sentence parsing into the initial choice of the sentence structure to be constructed, the assembly of this structure, the checking of the compatibility of the structure with the syntactic rules and, if necessary, the revision of the structure. This serial view with discrete stages of structure selection first and reevaluation and correction second has been criticised by theorists emphasising more incremental (e.g., Altmann & Kamide, 1999; Kaiser & Trueswell, 2004; Levy, 2008; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) and parallel and integrative sentence processing (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994; Tabor & Tannenhaus, 1999). In such approaches, expectations about the upcoming events in a sentence play an important role (e.g., Levy, 2008). For example, Tabor and Tannenhaus (1999) point out that “syntactic processing is simultaneously affected by semantic, syntactic and discourse-based information” (p. 492). The argument, however, about whether the parser uses both syntactic and semantic information interactively (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994; Marslen-Wilson & Tyler, 1980) or whether syntactic information is processed before lexical-semantic information (e.g., Frazier & Fodor, 1978; Friederici, Gunter, Hahne, & Mauth, 2004) is not yet resolved.

Regarding the structures used by the parsing system, most theorists (e.g., Gilboy, Sopena, Clifton, & Frazier, 1995; Frazier & Rayner, 1982; Levy, 2008; Van Gompel, & Pickering, in press) have based their models on the tree diagram representation introduced by Chomsky (1965, 1981). There have also been attempts to represent the parsing system in computational constraint satisfaction models without an underlying tree diagram structure (MacDonald, Pearlmutter, & Seidenberg, 1994; McClelland, St. John, & Taraban, 1989; Trueswell & Tanenhaus, 1994). McClelland and colleagues (1989), for example, point out that constraint satisfaction processing can deal with problems faced by more traditional approaches. One of these problems is that in different sentences, the same words can have different meanings and the same roles can have different functions. Consequently, different inferences can be drawn from the diverse combinations of meanings and functions. According to McClelland and colleagues, traditional models either make an early commitment to a particular combination of meaning and function (which might need to be revised) or keep track of a large number of possible combinations (which would be computationally demanding). A constraint satisfaction model, however, "avoids combinatorial explosion by keeping multiple alternatives implicit in the single pattern of activity over the sentence gestalt" (p. 316).

In sum, in order to comprehend a sentence, readers must process information on many different levels. To turn a sequence of features, letters and words into a meaningful conceptual representation, multiple steps must be carried out – some in sequence, some in parallel, some bottom-up, some top-down. If successful, these processes lead to an integrated sentence representation.

## 2.2 Representation of gender-relevant nouns

The target words in my experiments were nouns referring to gender-stereotypically female and male occupations (e.g., secretary, electrician). I therefore provide a short overview of how gender is processed in the English language. Although English does not have grammatical gender, it contains some gender-specified words. These are linguistically (semantically) defined as female and male in accordance with the natural gender of their referents (e.g., king/queen). There are only a few words where gender is also syntactically specified (e.g., actor/actress; waiter/waitress). Apart from gender-specified words, there are also a number of words with gender stereotypes attached to them. Certain occupations, for example, although not definitionally confined to one natural gender, are stereotypically expected to be held by women (e.g., secretary, babysitter) or men (e.g., electrician, mechanic).

For the purpose of this thesis, it is important to define whether gender stereotypicality is lexically or conceptually represented. Several researchers using stereotype-relevant occupation labels suggest that stereotypicality is part of a word's lexical representation. Duffy and Keir (2004), for example, investigated the influence of discourse context on the processing of stereotype-relevant words. In Experiment 1, they used anaphor[2] sentences with stereotypical occupation labels and matching or mismatching reflexive pronouns (e.g., "The electrician taught himself/herself…"). They tracked the reader's eye-movements and found longer viewing times in mismatching than matching sentences. Duffy and Keir used the lexical reinterpretation model as theoretical framework (Hess, Foss, & Caroll, 1995). They regarded the gender information connected to the occupation labels as part of the lexical representation, which can, however, be reinterpreted in a conflicting discourse context.

Sturt (2003) used similar anaphor sentences containing stereotype-relevant information in his second experiment (e.g., "The surgeon who treated Jonathan had pricked himself/herself with a used syringe needle"). He, too, measured participants' eye-movements in order to assess the time-course of anaphor resolution, reflected in the viewing times on the pronoun. Sturt found that his participants looked longer at pronouns that mismatched than at pronouns that matched the stereotype. These differences were already found in the earliest viewing-time measure (first fixation duration), which led Sturt to conclude that stereotypicality might be lexically marked.

[2] The term *anaphor* refers to words (typically pronouns) that are used to refer to an earlier part of a sentence (the referent or antecedent).

Other research, however, suggests that stereotypicality is part of a word’s conceptual representation. Carreiras, Garnham, Oakhill and Cain (1996), for example, measured self-paced reading time in response to texts like “The electrician examined the light fitting. He/She needed a special attachment to fix it”. They found longer reading times for stereotype-mismatching pronouns in the second sentence. Their account for this finding is that when a word like “electrician” is encountered, the gender is inferred from the prior knowledge of the stereotype and incorporated into the mental model of the text. When the gender is later made explicit as stereotype-mismatching, the mental model must be updated, which requires extra processing time. Gender information is interpreted as part of the conceptual rather than the lexical representation of the word. Kreiner, Sturt and Garrod (2008) argue that the only way to distinguish between the lexical and conceptual view is to compare the processing of definitional and stereotypical gender: If gender stereotypicality is part of the lexical representation of a word, it should be processed in much the same way as a gender-defined word. Kreiner and colleagues (2008) used anaphoric sentences like “Yesterday the king/minister left London after reminding himself/herself about the letter” and cataphoric[3] sentences like “After reminding himself/herself about the letter, the king/minister immediately went to the meeting at the office”. Eye-movement recordings revealed a mismatch-effect in the anaphor sentences for both definitional and stereotypical gender. In the cataphoric sentences, however, a mismatch effect was only observed for definitional gender, suggesting that stereotypical gender is not represented lexically. Similar differences between definitional and stereotypical gender have been found in other studies. Osterhout, Bersick and McLaughlin (1997) measured ERPs while participants

[3] The term *cataphor* refers to words that are used to refer to a later part of a sentence.

read anaphor sentences containing definitionally or stereotypically female and male agents and matching or mismatching reflexive pronouns (e.g., “The capable girl scout built herself/himself a fire”; “The popular babysitter found herself/himself overcommitted on Fridays”). Although mismatching pronouns in both sentence types elicited a P600 effect[4] — an effect similar to ones observed in response to syntactic anomalies (e.g., Hagoort, Brown, & Groothusen, 1993; Osterhout & Mobley, 1995) — it was larger for the definitional gender than the stereotypical gender. The authors point out that the difference between the two noun types is that “violations of gender definitions result in an unavoidable ungrammaticality, whereas violations of stereotypes force the less preferred gender assignment onto the antecedent noun” (p. 281). The explanation that stereotypical gender — in contrast to definitional gender — has to be inferred and can be reevaluated is consistent with the view of gender stereotypicality being conceptually rather than semantic-lexically represented.

Banaji and Hardin (1996) presented participants with either definitionally or stereotypically female and male prime words (e.g., “father”, “mother”, “doctor”, “nurse”) followed by pronouns (e.g., “he”, “she”). They found a gender-priming effect both in Experiment 1 where a gender decision about the pronoun was required and in Experiment 2 where a lexical decision was required. The effect in Experiment 1 was, however, significantly larger for definitional than stereotypical primes, and the effect in Experiment 2 was only reliable for definitional primes. Banaji and Hardin concluded: “This difference reflects the differential strength of the two types of primes in evoking gender. Words that are exclusively reserved to denote gender will

[4] The P600 is “A large positive wave with an onset at about 500 msec and a duration of several hundred milliseconds” (Osterhout, Bersick, & McLaughlin, 1997).

produce stronger priming than words that connote gender" (p. 140). This view is, again, consistent with a conceptual representation of stereotypical gender.

Given the evidence that gender-stereotyped nouns behave differently from lexically-semantically gender-specified nouns, I, too, adopt the conceptual approach.

Stereotype representation has also been considered outside of psycholinguistics, in social psychology. In social psychological stereotype research, representational models are often adopted implicitly, rather than being investigated in their own right (Hilton & von Hippel, 1996), but a few models have been specified.

Some researchers have assumed abstracted stereotype representations like schemas and prototypes (e.g., Brewer, Dull, & Lui, 1981; Hashtroudi, Mutter, Cole, & Green, 1984; Hilton & von Hippel, 1996). Others have challenged or expanded the prototype assumption with exemplar-based or mixed models (e.g., Linville, Fischer, & Salovey, 1989; Mullen & Johnson, 1995; Sherman, 1996; Smith & Zárate, 1992). Other ways of modelling stereotype representation were within associative (e.g., Devine, 1989) or connectionist networks (e.g., Smith & DeCoster, 1998). The ways in which these models differ in respect to how stereotypes are represented and organised is important because different representational assumptions can lead to different empirical predictions and explanations. Smith (1998), however, suggests that rather than considering the different model types as competitors, it might be more beneficial to detect how they can complement each other, "as having distinct (though occasionally overlapping) domains of applicability" (p. 429). Following Smith's suggestion to incorporate the advantages of different models described in the literature into hybrid models, the working model for this thesis incorporates elements of associative network models, schema or prototype models and exemplar models (see section 3).

## 2.3 Pronoun resolution

In my sentence stimuli, the gender–stereotype match or mismatch takes place when readers encounter the reflexive pronouns *herself* or *himself.* I therefore give a brief overview of pronoun processing.

In conceptual terms, a pronoun only refers to a referent. The process of finding the appropriate referent within a text is called pronoun resolution or binding. Many contemporary studies of binding (e.g., Kaiser, Runner, Sussman, & Tannenhaus, 2009; Sturt, 2003) are based on Chomsky's (1981) binding theory. Principle A of this theory details the syntactical constraints for binding anaphors such as reflexive pronouns. It states that "an anaphor is bound in its governing category" (p. 188), which, applied to the present context, means that a reflexive pronoun must be bound to an antecedent within the same sentence.

Generally, pronouns in English are gender-specific because they refer to the natural gender of their referents. In addition, however, there is ample evidence that pronoun comprehension is influenced by gender stereotypes associated with their antecedents. Carreiras, Garnham, Oakhill and Cain (1996, Experiment 1), for example, presented participants with sentence pairs. The first sentence contained a stereotypically female or male occupation label and the second sentence contained a matching or mismatching pronoun. Carreiras and colleagues (1996) found a mismatch effect on the reading time of mismatching second sentences. Kennison and Trofe (2003) also found a mismatch effect, using similar materials to Carreiras and colleagues, but measuring self-paced phrase-by-phrase moving window reading time. The mismatch effect could in this study be linked to the pronoun region because it was presented in a window on its own. Osterhout, Bersick and McLaughlin (1997) measured ERPs during word-by-word reading of sentence stimuli like "The popular babysitter found herself/himself overcommitted on Fridays" and found a mismatch effect on the pronoun for sentences with gender-stereotyped words. Finally, Banaji and Hardin (1994) found that participants recognised and categorised pronouns faster as female or male when they agreed with the stereotypical gender of prime words (e.g., "nurse", "doctor").

As I have mentioned, I assume that stereotypical gender is represented conceptually rather than lexically. I therefore also suggest that the mismatch effect arises on the conceptual level rather than as a result of a "clash of two sets of lexical features" as suggested by Sturt (2003, p. 560). In my view, it arises when the conceptual representation of a stereotypical noun has to be updated after an encounter with a mismatching pronoun.

# 3. A working model of conceptual stereotype representation

I assume that stereotypical gender is represented on the conceptual level and that the mismatch effect between gender-stereotypical nouns and reflexive pronouns arises on the conceptual level as well. I will outline a working model of conceptual representation, which will serve as a framework for the interpretation of my empirical results. Its scope will be limited to gender-stereotype representation and processing. The phenomena it will be able to model include the online processing effects of and memory performance for stereotype-relevant information with and without additional cognitive load during encoding (Experiment 1), the effects of episodic representation, constructed during reading stereotype-relevant information on further linguistic processing (Experiments 2 and 3) and the effects of cross-modal priming from written stereotype-relevant information to pictures (Experiments 4 to 7). Here I introduce the basic characteristics of the model and will come back to its interpretative functions when discussing my empirical results.

The working model comprises episodic exemplar and semantic prototype representations organised in the nodes and links of an associative network. As hybrid model, it benefits from the advantages of the different representational approaches (Smith, 1998). Episodic representations can generally be described as event memories and memories of the personally experienced past and will refer here to the representation of the information within a particular sentence. One characteristic of such episodic representations is that they can be consciously recalled. Semantic representations can generally be described as generic and world-knowledge representations and will refer here to stereotypical prototype knowledge (for an overview of the episodic/semantic distinction see Tulving, 1972, as cited in Tulving & Thomson, 1973, p. 354).

Concepts (e.g., secretary) and conceptual features (e.g., female) are presented as nodes within the network. The prototype and exemplar representations fulfil different functions within the network, yet are interdependent. The prototypes represent pre-experimentally acquired knowledge in form of the gender stereotypes associated with certain occupations. That means that within the network, secretaries, for example, are generally assumed to be female and mechanics to be male. Exemplar representations are formed by linking different nodes when new stereotype-relevant information is encountered. These exemplars can be stereotype-matching (e.g., female secretary) or stereotype-mismatching (e.g., male secretary). The prototype representations (e.g., of a woman or a secretary) are assumed to be made up of an abstraction of multiple exemplars (see Clark, 1993). They can change or be updated through a statistical learning mechanism, taking into account the information about new exemplars. This learning mechanism is assumed to be very slow so that individual exemplars do not change the semantic representation too much.

The prototype representations fulfil interpretative functions when new information is encountered and when a particular event is attempted to be remembered. When new prototype-relevant information is encountered (e.g., the word *secretary*), connected features are activated, (e.g., *female*), resulting in an expectation (e.g., that the secretary will be female). When a specific processing event is attempted to be remembered, the episodic exemplar representation can sometimes not be reconstructed. In this case, the prototype representation offers a pattern completion function by providing general, abstracted information as guessing and reconstruction aid.

Generally, the ease with which an episodic representation can be reconstructed is assumed to be reflected in the processing effort devoted to it during its formation and thus the strength and stability of the representation. Carlston and Smith (1996, as cited in Smith, 1998) termed this the processing by-product principle. The effort that can be allocated to the processing is restricted by the cognitive resources available.

## 4. Overview of the thesis

Using the outlined working model as a framework for the interpretation of my empirical results, I report in the second chapter Experiment 1, investigating the effects of cognitive load on online processing of and memory for stereotype-relevant information. In Chapter 3, I move on to reporting Experiments 2 and 3, examining the effects of episodic representations constructed during reading stereotype-relevant information on subsequent linguistic processing. In Chapter 4, I report Experiments 4 to 7, looking into the effects of stereotype-relevant linguistic context on subsequent non-linguistic, pictorial processing. In Chapter 5, I summarise my findings, integrate them into the working model and discuss their limitations as well as their psycholinguistic and wider social relevance.

# Chapter 2
# Effects of cognitive load on online processing of and memory for stereotype-relevant information

# 5. Experiment 1

## 5.1 Overview and goals

In this chapter, I sought to replicate the mismatch effect and to assess memory for stereotype-relevant information. I further examined the effects of cognitive load on online processing and memory. Here, as in previous studies (e.g., Carreiras, Garnham, Oakhill, & Cain, 1996; Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Osterhout, Bersick, & McLaughlin, 1997; Sturt, 2003), the stimuli were sentences that included stereotype-matching or -mismatching occupation–pronoun combinations (e.g., “Last week the *secretary* familiarised *herself/himself* with the new photocopier”). During reading, the participants’ eye movements were recorded. Half of the participants carried out a concurrent cognitive load task during reading (*load condition* hereafter), whereas the other half did the task without additional cognitive load (*no-load* condition hereafter). I measured viewing times for the agent region (e.g., *secretary*), the pronoun region (e.g., *herself/himself*) and the region immediately following the pronoun region (e.g., *with*). After the reading task, participants’ recall of the agents’ gender was assessed (e.g., “Was the secretary male/female?”).

One goal of the experiment was to replicate the finding that readers look longer at pronouns that mismatch rather than match the stereotypical gender of the agent (Duffy & Keir, 2004; Kreiner et al., 2008; Sturt, 2003). It was important to establish the basic effect as a foundation for studying the relationship of online processing and memory and for examining the effect of cognitive load on the mismatch effect.

Another goal of the study was to assess participants’ memory for stereotype-relevant information. This was important because examining what participants remembered of the stereotype-relevant information can give an insight into the representations they construct during online processing. Differences in the online processing times found in previous studies (Duffy & Keir, 2004; Kreiner et al., 2008; Sturt, 2003) only indicated that a mismatch was detected. The data from these studies do not indicate whether the increased processing time for stereotype-mismatching versus -matching information supported the generation and maintenance of a lasting mental representation of the agent that includes the gender information provided by the pronoun, or whether it reflected the construction of a temporary representation in order to comprehend the sentence.

A further goal of the study was to test whether online attention allocation to and subsequent memory for stereotype-relevant information are affected by cognitive load. This question is an important one because it sheds light on whether people automatically[5] detect and resolve a mismatch when reading about a disconfirming member of a stereotyped group, even when cognitively busy, or whether mismatch resolution is a capacity-demanding process. If mismatch resolution is demanding, then mismatches might remain unresolved under cognitive load. Whether mismatch detection and resolution are automatic or capacity-demanding processes has important consequences for the change of representations of stereotyped groups, because the likelihood of a representation update would be crucially diminished when mismatch detection and resolution could only take place when attention is undivided..

[5] In line with Bargh's (1994) claim for specific definitions of which qualities of automaticity are being investigated, I define, for the purpose of this thesis the terms capacity-demanding and automatic dichotomously. If a task can only be carried out when enough working memory capacity is available, it is defined as capacity-demanding, otherwise as automatic.

## 5.2 Hypotheses

### Measuring the online processing of stereotype-relevant information with or without cognitive load

The first goal of the study was to replicate the mismatch effect (Duffy & Keir, 2004; Kreiner et al., 2008; Sturt, 2003). I expected to replicate previous findings that the integration of mismatching pronouns is more time-consuming than the integration of matching pronouns, reflected in longer viewing times and more regressions[6] back into earlier regions of the sentence. The expectations for specific sentence regions are detailed in the method section 5.3.

The second goal of the study was to investigate the processing demands of mismatch resolution. It is not clear from previous findings (Duffy & Keir, 2004; Kreiner et al., 2008; Sturt, 2003) whether the process of resolving a gender mismatch is an automatic or capacity-demanding process. The mismatch only has to be resolved when participants infer the stereotypical gender in the first place. McKoon and Ratcliff (1992) studied the process of establishing the connection between an anaphor and its referent and pointed out that inferences are – in absence of a specific goal – only made when explicit detail is given in the text and in cases of “information about potential referents being quickly available” (p. 444). This *Minimalist Hypothesis* therefore suggests that drawing further-reaching inferences in anaphor resolution is a strategic and effortful rather than automatic process. Carreiras, Garnham, Oakhill and Cain (1996), however, argued that McKoon and Ratcliff did not give a definition of what type of information and knowledge is easily available. It is not clear whether stereotype-relevant information falls under this description. The results by Reynolds, Garnham and Oakhill (2006) suggest that inferences, in particular about stereotypical

[6] Regressions are defined as right-to-left movements to previously read words (Rayner, 1998).

gender, are made immediately and at least partly automatically. They pointed out that participants in their study took longer to read a final sentence in a passage when the agent's gender mismatched (versus matched) the stereotypical gender introduced earlier in the passage. Participants also found the passage more difficult to comprehend. Reynolds and colleagues concluded that the participants must have immediately and automatically drawn an inference when first processing the stereotype-relevant agent. None of these studies, however, tested whether inferring stereotypical gender and resolving a mismatch is dependent on readers having plentiful cognitive capacities.

To study this, I manipulated the cognitive load in the reading task. Half of the participants read the sentences with additional cognitive load and half of the participants without cognitive load. The load and no-load conditions were originally carried out as separate experiments, but will be reported together as Experiment 1 with cognitive load as between-participants variable. For the cognitive load manipulation, I chose a working memory load task. Working memory capacity has been proposed to constrain language comprehension (Just & Carpenter, 1992). Because the task was not to interfere with the linguistic processing of the sentences, I used a continuous non-linguistic 5-digit retention task rather than a discontinuous task like tone or probe monitoring.

There have been some social psychological studies into the effects of cognitive load on the processing of stereotype-relevant information. Sherman, Lee, Bessenoff and Frost (1998), for example, studied the effect of, cognitive load on attention allocation to stereotype-relevant information[7], reflected in sentence-reading time. In Experiment

[7] Eight-digit number retention task during the reading of 30 statements

1, participants formed an impression of a skinhead or a priest. Reading times per sentence were measured using a self-paced reading task (one sentence at a time). The sentences contained behaviours that were consistent or inconsistent with the skinhead or priest stereotype. Sherman and colleagues found that participants devoted similar amounts of time to both sentence types when processing capacity was high. Under cognitive load, however, they attended more to inconsistent than consistent sentences. In their third experiment, Sherman et al. used the same impression formation task as before, but with two sentences, one stereotype-consistent and the other -inconsistent. Because presentation time was limited to four seconds, participants had to choose which sentence to attend to. Processing capacity was, as before, high or low. Recognition accuracy as measure for encoding effort was the same for inconsistent and consistent items under high capacity conditions and better for inconsistent than consistent items under low capacity. Sherman and colleagues explained these findings within the Encoding Flexibility Model (see also Sherman, Conrey, & Groom, 2004; Sherman & Frost, 2000). They argued that stereotypes provide an efficient tool to extract both stereotype-matching and -mismatching information, even when cognitive resources are low. According to Sherman and colleagues, stereotypes provide the conceptual fluency for the gist of stereotype-matching information to be extracted in a capacity-saving way. Being able to rely on stereotypes for the processing of stereotype-matching information in turn allows the perceiver to devote more attention to difficult-to-comprehend stereotype-mismatching information, especially when resources are depleted. Overall, the results in the study by Sherman and colleagues indicate that in impression formation tasks, the processing effort for stereotype-mismatching information can increase under cognitive load.

My task, however, did not have an online impression formation goal. Evidence elsewhere has shown that an impression formation goal can change the way people process information. Hastie and Park (1986) asked their participants either before or after presenting them with a conversation between a target person and another man to form an impression of the person (online versus memory-based impression formation, respectively). The results revealed a correlation between information recall and impression judgements for participants in the memory-based condition but not the online condition. Given the difference an impression formation goal can make, the results of Sherman and colleagues (1998) might have only limited predictive power for the present study. As a result, my hypotheses were not derived exclusively from their results.

I expected that cognitive load would have an effect on overall reading times, as well as reading times for the specific stereotype-relevant regions. Belke (2008) found the 5-digit retention task to slow down word and picture naming latencies. It is therefore possible that it could slow down the entire sentence reading process too due to the additional cognitive demands. This would be reflected in longer overall viewing times and possibly more regressions back into earlier regions of the sentence.

I expected further that if the recognition and integration of stereotype-mismatching information are automatic processes, then the processing differences between matching and mismatching information would be similar in the load and no-load conditions. If they are, however, capacity demanding processes, I expected the online processing differences between the stereotype-matching and mismatching information to decrease or disappear in the load condition.

The viewing-time measures reflecting early and late processing could have diverged or converged. If the mismatching information was not even noticed under cognitive load, I expected the viewing time differences reflecting both early and late processing to decrease or disappear. If, however, the mismatch was noticed but not integrated under cognitive load, I expected the viewing time differences reflecting late but not early processing to decrease or disappear.

### Measuring memory for stereotype-relevant information after encoding with or without cognitive load

A memory questionnaire for stereotype-relevant information was included at the end of the reading task. In a cued-recall questionnaire, participants reported whether the occupations mentioned in the sentences had been held by women or men. By applying signal detection theory, I determined not only memory sensitivity, but also the size and direction of any response bias, depending on the cognitive load during encoding. The importance of distinguishing between memory sensitivity and bias has been pointed out by Stangor and McMillan (1992). Within their meta-analyses of memory for expectancy-congruent and expectancy-incongruent information, differential effects were found for sensitivity and bias. Memory sensitivity favoured expectancy-incongruent information, whereas response bias favoured expectancy-congruent information. Memory sensitivity is a measure of the correspondence between the presented information and participants' memory. Memory bias is a measure of the tendency to respond in a stereotype-matching or -mismatching way, regardless of the information presented.

Previous studies using eye tracking during reading of stereotype-relevant information (Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Sturt, 2003) have not assessed memory after the reading task and it is therefore not clear to what extent detected mismatches were resolved and whether correct representations were formed – particularly of the mismatching information.

There has, however, been some social psychological research into memory for stereotype-relevant information. Most of these studies have focused on whether stereotype-matching or -mismatching information is remembered better. A memory advantage for matching information was suggested by the findings by Rothbart, Evans and Fulero (1979), ostensibly because social schemas filter out mismatching information. Other studies suggested a memory advantage for mismatching information (Bargh & Thein, 1985; Hastie & Kumar, 1979; Macrae, Hewstone & Griffith, 1993; Sherman & Frost, 2000). Hastie and Kumar (1979) offered as explanation for this result that the more informative an event is, the deeper it is processed. The deeper an event is processed, the more likely it is to be remembered later. Hastie and Kumar suggested that mismatching information is better remembered than matching information because the most informative events are novel and unexpected (mismatching) ones.

Whether expectancy-matching or -mismatching information is remembered better depends on many factors, as has been shown by the meta-analysis by Stangor and McMillan (1992). They found, for example, differential effects for recall and recognition measures, with recall and recognition sensitivity tending to favour expectancy-mismatching information and bias-uncorrected recognition measures and response bias tending to favour expectancy-matching information. They also emphasised the role of moderator variables on memory performance - for example, processing goals, the strength of the expectation, the complexity of the presented information, processing time, whether memory for groups or individuals was tested and whether the expectations were pre-experimentally existing or experimentally created. Further influences have been suggested to come from moderator variables like order of stimulus presentation, proportion of matching and mismatching stimuli and stimulus exposure time (Rojahn & Pettigrew, 1992).

The evidence for the effects of cognitive load on memory for stereotype-relevant information is also not clear-cut. Macrae, Hewstone and Griffiths (1993) and Bargh and Thein (1985) found a recall advantage for mismatching over matching information when processing capacity was high. When processing capacity was low, however, this advantage disappeared. Macrae and colleagues suggested that this pattern arose because when enough cognitive resources were available, participants processed mismatching information in greater depth in order to resolve the inconsistencies. When cognitive resources were low, however, this processing advantage disappeared. Sherman and Frost (2000), however, found a recognition advantage for mismatching information under cognitive load, but a recall advantage for matching information. Using the Encoding Flexibility Model, they argued that because the gist of matching information can easily be inferred using the stereotype, perceivers direct more encoding effort to mismatching information when processing capacities are restricted. This encoding effort leads to the recognition advantage for mismatching information, whereas the retrieval facilitation provided by stereotypes lead to a recall advantage for matching information. This dissociation for different memory measures was only one of the points indicated by Stangor and McMillan (1992) and Rojahn and Pettigrew (1992) to affect memory results for schema and stereotype-matching and -mismatching information. Therefore, the results of studies with a different set-up can only give very limited indication of what to expect in the present study.

The utility of past research in memory for stereotype-matching and mismatching information was also limited by design differences between the present study and many previous studies. First, the current focus was not on comparing whether matching or mismatching information was remembered better, but to assess participants' overall memory sensitivity and bias. These measures can give an indication of whether correct representations were constructed during encoding. Second, impression formation tasks have typically been used with only a small number of specific targets (e.g., Hastie & Kumar, 1979; Macrae, Hewstone, & Griffiths, 1993; Rothbart, Evans, & Fulero, 1979; Sherman & Frost, 2000). Here, however, memory was tested for a larger number of unspecified targets that each appeared only once over the course of the reading task. In impression formation tasks, targets are generally described by several statements or traits. The time and effort spent on processing these targets presumably results in more in-depth representations than the ones formed in sentence reading tasks. I chose the present method over an impression formation design because I was interested in investigating the online attention allocation and the resulting representations of and memory for stereotype-relevant information rather than in the higher-level processes of forming an integrated impression of a particular person.

I expected the memory results to depend on the effect of cognitive load in the online reading task. In case the online recognition and integration of stereotype-mismatching information are automatic processes, resulting in similar online processing differences between matching and mismatching information in the load and no-load condition, I expected participants to be able to integrate the mismatching pronoun information in both load and no-load conditions of the reading task and to construct an appropriate representation of the agents. I therefore expected memory sensitivity to be above chance in both load and no-load conditions. I did not anticipate perfect memory for the sentence information, however, and expected participants to have a conservative memory bias, using their stereotypes as guessing and reconstruction cues when they could not remember an agent’s gender. I expected that if encoding information under cognitive load makes reconstructing mismatching information more difficult, the bias should increase in the load condition compared to the no-load condition.

In case the recognition and integration of stereotype-mismatching information are capacity demanding processes, resulting in a decrease or disappearance of the online processing differences between the stereotype-matching and mismatching information in the load condition, I expected participants to be able to integrate the mismatching pronoun information and to construct an appropriate representation of the agents only in the no-load condition , resulting in memory sensitivity to be decreased in the load condition compared to the no-load condition. In this case I also expected that reconstruction of mismatching information should be much harder in the load condition, resulting in an increased bias in the load compared to the no-load condition.

Participants read a large number of sentences before filling in the gender cued-recall questionnaire. To test whether there was sufficient memory overall to draw conclusions from the gender-memory questionnaire, a baseline sentence-memory test was included in the load condition[8], asking only for stereotype-irrelevant information. Participants read sentences with two possible endings: the original and a new, equally sensical, option (e.g., “Last week the secretary familiarised herself with *the new photocopier/the new software*.”) and reported which ending had been presented at

[8] The load and no-load conditions were originally conducted as separate experiments and the questionnaire was only included in the load experiment. In hindsight the baseline questionnaire should have been included in both experiments. However, as will be shown in the discussion, the time between encoding and cued recall did not affect memory sensitivity or bias, so it is reasonable to assume that the results of the baseline questionnaire should also hold for the no-load experiment.

encoding. If participants processed the information thoroughly during encoding, the results of this questionnaire should be well above chance.

## 5.3 Method

### Pretest: Stimulus selection

The occupation labels as stimuli for Experiment 1were pretested for their female/male typicality. Before moving on to the method of the main experiment in this chapter, the method and results of this pretest will be reported, followed by an overview of the use of eye tracking in investigating reading processes.

#### *Method*

Twenty-five female undergraduate students at the University of Birmingham took part in the pretest. All were native speakers of British English. Participants received course credits or money in exchange for their participation.

Participants rated 89 occupations for female and male typicality, responding to two questions for each: “Do you regard this occupation as typically female?” and “Do you regard this occupation as typically male?”. Responses were made along 7-point scales anchored by 1 (*not at all*) and 7 (*very much so*). The two scales were used to select occupations that were at the same time stereotypically female but not stereotypically male, and vice versa.

#### *Results*

In one-sample *t*-tests carried out separately for each item, the mean scores for female and male typicality were compared to the midpoint value of the scales (4). Twelve strongly stereotypically female and 12 strongly stereotypically male items were selected. Stereotypically female occupations were those with mean female typicality ratings greater than 6 and mean male typicality ratings lower than 2: *beautician, babysitter, midwife, florist, receptionist, secretary, cheerleader, childminder, housekeeper, fortune teller, typist* and *nanny*. Stereotypically male occupations were those with mean male typicality ratings greater than 6 and mean female typicality ratings lower than 2: *plumber, lorry driver, carpenter, bricklayer, locksmith, butcher, mechanic, taxi driver, pilot, construction worker, footballer* and *security guard*.

### Use of eye-tracking to investigate reading processes

Eye-movement measures have long been used in reading research to infer cognitive processes. It has been shown that they are closely related to moment-to-moment cognitive processes (e.g., Rayner, 1998; Reichle, Pollatsek, Fisher, & Rayner, 1998). In reading comprehension, the effects of low-level visual and medium-level cognitive variables (e.g., word frequency) on eye gaze control have been particularly well researched (e.g., Rayner, 1998). By contrast, higher level processes such as plausibility or reader's bias are explicitly excluded from current computation models of eye gaze control in reading (e.g., Engbert, Nuthmann, Richter, & Kliegl, 2005; Reichle et al., 1998). It has been shown, however, that eye-movements reflect attention allocation and effort of processing during reading stereotype-matching or mismatching information (e.g., Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Sturt, 2003). I therefore chose to use this method to determine how the online processing differed between stereotype-matching and mismatching information.

## Method of Experiment 1

### *Participants*

Sixty-one participants completed the experiment (46 female). The no-load condition included 29 participants (18 female; mean age 21.86 years, ranging from 18 to 40 years). The load condition included 32 participants (28 female; mean age 20.72 years, ranging from 17 to 33 years). All participants were undergraduate or postgraduate students at the University of Birmingham[9] with normal or corrected-to-normal vision and all were native speakers of British English. They received either course credits or money in exchange for their participation.

### *Apparatus*

I used the experimental software package SR Research Experiment Builder to create the experiment. The stimuli were presented on a 22 inch ViewSonic P225f monitor. Eye-movements were recorded using a video-based SMI EyeLink II head-mounted eye-tracking system with a data rate of 500 samples per second. For the eye-data analysis, EyeLink Data Viewer software was used. “Yes” and “no” responses were recorded with push-buttons on the EyeLink II response box.

### *Materials*

#### *Reading-task materials*

For the reading task, 24 sentences were constructed around the stereotypically female and male occupation labels selected in the pretest. Each of the sentences had two versions (see Appendix 1). In one version, the occupation label was combined with a stereotype-matching reflexive pronoun (e.g., “Last week the *secretary* familiarised

[9] One participant was tested during an internship.

*herself* with the new photocopier"; emphasis added) and in the other version, the occupation was combined with a stereotype-mismatching reflexive pronoun (e.g., "Last week the *secretary* familiarised *himself* with the new photocopier.").

Each sentence contained an *initial region* (e.g., *Last week*), followed by an *agent region* (e.g., *secretary*), a *verb region* (e.g., *familiarised*), a *pronoun region* (e.g., *herself/himself*), a *pronoun spill-over region* (e.g., *with*), and a *final region* (e.g., *the new photocopier*). The initial region contained at least two words and consisted of information that, according to English grammar, naturally precedes the subject of a sentence, for example, a locator of time (e.g., *Last week*, *In the evening*). The agent region contained the subject of the sentence. It consisted of a noun (e.g., *secretary*) or compound noun (e.g., *lorry driver*), specifying an occupational role. The verb region consisted of a transitive verb. The pronoun region consisted of a reflexive pronoun. The spill-over region consisted of the word following the pronoun if that word had four or more letters, or two words following the pronoun otherwise. This region was included because it has been shown that sometimes the processing of a word is not only reflected in the fixations on that word, but also influences the fixations on the following word (e.g., Duffy & Rayner, 1990; Rayner & Duffy, 1986). The final region contained at least one word. The spill-over and final regions contained information specifying the object or circumstances of the subject's action.

The sentences were displayed in Times New Roman Font, Size 18 and fitted on one line on the computer screen. One character had an average degree of visual angle of 0.35°.

In order to disguise the purpose of the reading task, 54 filler sentences were randomly mixed in with the experimental sentences (e.g., "Sammy first noticed the dragon fly when he woke up from his nap in the hammock."). The filler sentences served as targets for two other experiments unrelated to the present study and had a different linguistic structure than the experimental sentences.

To keep participants alert and assess their overall comprehension, half of all experimental and filler sentences were followed by simple *yes*/*no* comprehension questions (e.g., “Was Sammy sleeping in the hammock?”). Altogether 40 comprehension questions were presented. *Yes* and *no* push-button responses were recorded.

In the load condition, each sentence was preceded by a 5-digit number (*load number* hereafter). Half of the sentences were followed by the same number, and half by a different 5-digit number (*probe number* hereafter). Altogether 38 probe numbers were presented. The numbers did not include the digit 0, any immediately repeated digits (e.g., **11**516), or any repeated digit pairs (e.g., **16**5**16**). If the load and probe numbers were different, the changes were minimal in order to make recognition more difficult: Either one digit was replaced (e.g., load number 16582, probe number 16572) or two adjacent digits were exchanged (e.g., load number 16582, probe number 16852).

*Memory-tasks materials*

The gender cued-recall questionnaire (see Appendix 2) tested the accuracy of participants’ post-experimental representation of the stereotype-relevant information. It contained 24 items asking for the gender of the agents that had appeared in the experimental sentences (e.g., “Was the secretary male/female?”). Each response alternative had a tick box next to it. The items appeared in a different order from the one in the reading task.

In the load condition, an additional baseline-memory measure was included. This memory measure did not require the recall of any stereotype-relevant information.

The sentence-memory questionnaire contained the 24 experimental sentences with two alternative endings (see Appendix 3): the originally presented ending and an alternative ending similar in structure and content (e.g., "Last week the secretary familiarised herself with *the new photocopier/the new software*."). Each response alternative had a tick box next to it for participants to indicate which ending they believed to be the original one. The questionnaire had two versions that matched the agent-pronoun combinations of the sentences in the reading task.

### *Design*

The experiment was based on a 2 (occupation gender: female, male) x 2 (pronoun gender: female, male) x 2 (cognitive load: no load, load) mixed design with cognitive load as a between-participants factor.

The experiment included a no-load condition and a load condition. In the load condition, each sentence was preceded by a 5-digit load number and on half the trials followed by a 5-digit probe number. In half of the cases, the probe number was the same as the load number, and in half of the cases, it was different from the load number. Participants indicated their same/different judgements using push-button response. The remaining sentences were followed by a comprehension question.

Twelve of the 24 experimental sentences included a stereotypically female occupation label, and 12 included a stereotypically male occupation label. Six of each of the female and male sentences were combined with a stereotype-matching pronoun (match condition), and the other six sentences with a stereotype-mismatching pronoun (mismatch condition).

The experiment was tested in two versions between participants: All sentences that were stereotype-matching in Version 1 (e.g., "Last week the *secretary* familiarised *herself* with the new photocopier.") were stereotype-mismatching in Version 2 (e.g., "Last week the *secretary* familiarised *himself* with the new photocopier."), and vice versa.

Each version was presented in two different orders across participants to control for order and fatigue effects. In the no-load condition, the sentences were presented in four blocks; half of the participants saw first Blocks 1 and 2 followed by Blocks 3 and 4, and the remaining participants saw first Blocks 3 and 4, followed by Blocks 1 and 2. In the load condition, the experiment lasted longer due to the additional load task and was therefore split into six blocks. Half of the participants saw first Blocks 1 to 3 and then Blocks 4 to 6; the other half saw first Blocks 4 to 6 and then Blocks 1 to 3.

To encourage reading for comprehension, participants responded to comprehension questions after half the sentences. The questions required a "yes" or "no" push-button response. In the load condition, these questions occurred on trials where there was no probe number (i.e., each sentence was preceded by a load number and followed by either a probe number or a comprehension question). This design was chosen to ensure that participants would attend to both the sentence reading and the number-retention task.

In summary, trials in the no-load condition of the reading task consisted of either a sentence or a sentence and a comprehension question, and trials in the load condition consisted of a load number, a sentence, and either a probe number or a comprehension question.

After the reading task, participants in both no-load and load conditions completed a gender cued-recall questionnaire. For each sentence they had read in the reading task they indicated whether the agent had been female or male by ticking the appropriate box.

In the load condition, participants also filled in a sentence-memory questionnaire as memory-baseline measure that did not require the recall of any stereotype-relevant information. The questionnaire included all experimental sentence stimuli, each of which was presented with two alternative endings (in a different order from the reading task). Participants indicated which ending they thought they had read by ticking the appropriate box. The questionnaire had two versions of agent-pronoun combinations to match the agent-pronoun combination in the reading task versions.

### *Procedure*

Participants completed a consent form and a short questionnaire specifying age, gender, and first language. Then they read instructions describing their tasks (see Appendix 4 for the instructions for the load condition. The instructions for the no-load condition were adapted accordingly).

Participants moved through the trials (i.e., sentences, comprehension questions and, in the load condition, numbers) at their own pace. They sat at a distance of 70 cm from the computer monitor. A height-adjustable chin rest was used to reduce head-movements.

The eye-tracker was calibrated and validated. The calibration and validation results are used by the eye-tracker software to automatically calculate gaze positions during the experiment. During calibration, participants fixated on dots on the screen that appeared in random order in predefined positions within a three by three grid. The calibration was successful when for each dot position a fixation with no more than 1.5 degree deviance was recorded. Otherwise it was repeated until this criterion was met. After the calibration, participants repeated the task of fixating on the dots. These fixation data were then compared to the calibration data. Validation was successful when there was less than one degree difference on average or less than 1.5 degree maximal deviation. Otherwise both calibration and validation were repeated. For recalibration the eye cameras could be readjusted.

After calibration and validation, a drift correction followed, correcting for the natural variance in the participants' pupil positions. Participants fixated here on a dot in the middle of the screen. The experimenter could see the dot and the pupils' position on the experimenter monitor and accepted the drift correction when the participant's pupils overlapped the dot. The eye-tracking system was now prepared to record the eye movements accurately and the experiment could start. At the beginning of each trial, a drift correction was repeated at the position the first word of the sentence was about to appear. This procedure allowed the experimenter to see whether the headband had shifted due to head movement (e.g., during the breaks) and when recalibration was necessary.

After the reading task, participants completed a gender cued-recall questionnaire and, in the load condition, a sentence-memory questionnaire.

The experiment lasted approximately 30 minutes in the no-load condition and 45 minutes in the load condition. After completion, participants were debriefed and received course credits or money.

### ***Statistical analyses***

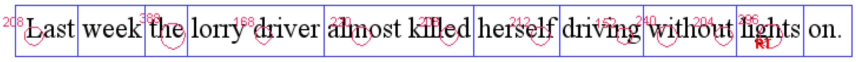
For all eye-movement measures, I conducted analyses of variance, reporting $F_1$ using participants as random variable and $F_2$ using items as a random variable (Clark, 1973). For the cognitive load manipulation measure, comprehension questions and the sentence-memory questionnaire, I conducted one-sample *t*-tests. For the analyses of the gender cued-recall questionnaire, I applied Signal Detection Theory and performed log-linear analyses.

***Analysis of eye-movements***

I used the graphical EyeLink II software EyeLink Data Viewer for the analyses of eye-movements. It showed the sentences a participant has seen, divided up in interest areas for each word or word combination, and fixations upon it as superimposed circles (see Figure 1 for a screenshot). For each fixation, the average position, the in-time and out-time as well as the duration were available. Consecutive fixations below 80 msec and within one character of each other were set to automatically merge to one fixation (see Kreiner, Sturt, & Garrod, 2008). Rayner and Pollatsek (1989) have argued that during fixations as short as this, participants cannot extract much meaning. The data from the right eye were analysed.

Participants were instructed to look at the fixation dot at the beginning of each trial. Sometimes the first fixation of a trial was located slightly above or below this fixation dot. Typically, this was seen on several successive trials. These drifts were corrected by manually moving the positions of all fixations of that trial up or down, so the first fixation would align with the fixation mark.

Figure 1: Screenshot of a sentence with superimposed fixations as displayed by the analysis software EyeLink Data Viewer



Last week the lorry driver almost killed herself driving without lights on.

*Choice of dependent variables: eye-movement measures*

A variety of eye-movement measures were used to gain insight in the different processes happening during reading comprehension. The measures can be clustered into early and late viewing-time measures as well as regression-proportion measures.

The division between early and late measures refers to the fact that inferences can be drawn about early versus late processes by looking at the pattern of results obtained for different measures. It is important to note that the measures are cumulative and therefore not independent of each other. For a diagram explaining the viewing-time measures with an example sentence, see Figure 2 below.

The following viewing-time measures were taken.

*First Fixation Duration* is defined as the duration of the first fixation falling in an interest region. It was the earliest measure, reflecting early cognitive processes including word recognition and early lexical access (Rayner, Juhasz, & Pollatsek, 2005).

*First-Pass Duration* is defined as the time interval between the onset of the first fixation and the offset of the last fixation on an interest region before the shift of gaze to the right or left. Both First Fixation and First-Pass Duration are sensitive to cognitive variables (e.g., word frequency, predictability) and some syntactic violations (Rayner, 1998; Pickering, Frisson, McElree, & Traxler, 2004). In the following, they will be referred to as early viewing-time measures.

*Selective Regression-Path Duration* is defined as the durations of all first-pass fixations on an interest region plus fixations made on that interest region after a leftwards regression. This is a measure of the time spent on a word until it is comprehend well enough to move on[10].

[10] Another frequently used measure is Regression-Path Duration, including all fixations made on a region and during regressions to the left until the eyes leave the region to the right (e.g., Duffy & Keir, 2004; Sturt, 2003). It has not been included here because it contains all rereading fixations on earlier parts of the sentence, which can make its interpretation difficult (see Pickering, Frisson, McElree, & Traxler, 2004). Nevertheless, these analyses were carried out for the present experiment and yielded no significant effects.

*Total Reading Time* is defined as the sum of all fixation times on an interest region and is an even later measure than Selective Regression-Path Duration.

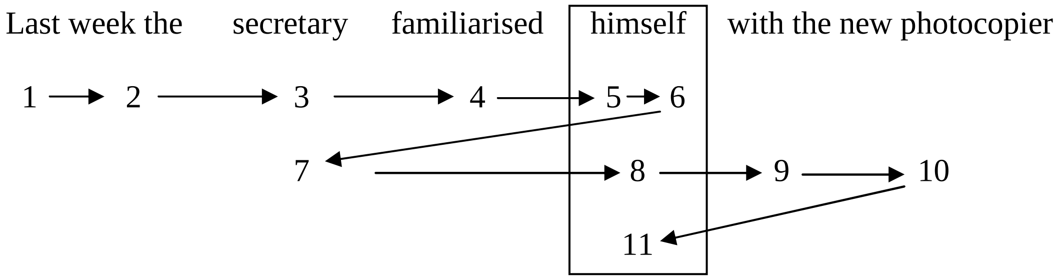
*Sentence Reading Time* is defined as the total reading time on the sentence region and *Sentence Fixation Count* as the total number of fixations made on the sentence region. These measures were included to capture processes happening relatively late in comprehension.

The following regression-proportion measures were measured.

*Regression Out* is defined as the proportion of times relative to the number of valid trials where at least one regression was made leftwards out of an interest region before moving on to the right. As only first-pass regressions are considered, Regression Out is regarded an early measure of word processing (see Pickering et al., 2004).

*Regression In* is defined as the proportion of times in which at least one regression has been made into the interest region from later parts of the sentence. It is considered to reflect later processing.

Figure 2: Example of viewing-time measures (on pronoun region)



**Viewing-time measures for the pronoun and sentence region:**

| **Measure** | **Example fixations** | **Processes reflected** |
|---|---|---|
| First Fixation duration | 5 | Only early processes |
| First-Pass Duration | 5 + 6 | Only early processes |
| Selective Regression Pass Duration | 5 + 6 + 8 | Early + later processes |
| Total Reading Time | 5 + 6 + 8 + 11 | Early + later processes |
| Sentence Reading Time | 1 + 2 + .... + 11 | Early + late processes |

*Choice of regions of interest*

The pronoun region, the pronoun spill-over region and the agent region were selected as regions of interest for the analyses of the eye data based on the choices of interest regions in earlier studies (e.g., Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Sturt, 2003). The entire sentence region was selected based on the theoretical consideration that it might capture late integrative comprehension processes that might not be seen on the single-word level. I will here specify my expectations for each region.

For the *pronoun region***,** I expected early as well was as later mismatch effects[11]. This is the main region of interest as here the occupation–gender stereotype is either matched or mismatched.

For the *pronoun spill-over region***,** I expected early mismatch effects, depending on how much of the processing of the pronoun continued after the eyes move on.

For the *agent region***,** I expected later but not early effects and that participants would look back to this region more frequently after encountering a mismatching than a matching pronoun. I expected that integrating the agent with the pronoun would result in late mismatch effects.

For the *entire sentence region***,** I expected the overall viewing time to be longer for mismatching than matching sentences. Overall viewing time reflects late comprehension and context integration processes, which might or might not be reflected on the single-word level.

The Data viewer software automatically created narrow rectangular boxes around each word (see Figure 1). I used these boxes in the interest regions analyses for the pronoun, pronoun spill-over and agent regions. I manually created combined interest regions for agent regions comprising compound words (e.g., lorry driver) or spill-over regions with two words by merging the boxes. For the entire sentence region, I used the sum of fixations over all individual interest regions.

### *Exclusion criteria*

A region of interest was excluded from the analyses if it had been skipped during first-pass reading. The reason for this was that it is difficult to interpret fixations on an

[11] Due to the lack of consensus in the literature, the distinction between early and late effects refers here only to chronological processing time and does not imply any assumptions about the underlying cognitive processes.

interest region after words further to the right have been read: The interest region might have already been viewed parafoveally and the fixation times might be influenced by the sentence context. An interest region was also excluded from the analyses when a participant had blinked while gazing at it.

Following common practice in eye-movement research, interest regions with outlying First Fixation Durations were excluded from all analyses. As the lower limit, I chose First Fixation Durations smaller than 80 msec (e.g., White, 2008). As the upper limit, I chose 600 msec. I chose a lower limit here than used in other studies (e.g., Niswander, Pollatsek, & Rayner, 2000) because very few First Fixation Durations exceeded even the 600 msec boundary. For the entire sentence region analyses, trials were excluded when Sentence Reading Times were greater than 8 seconds.

For each region of interest, the exclusion criteria were applied separately. That means that if, for example, a blink occurred on the agent region of a trial, this region, but not the pronoun and spill-over regions, were excluded from the analyses of this trial. This approach was chosen over one of excluding the entire trial if any of the interest regions were excluded in order to minimise the data loss. I assumed that the processing of one region could be assessed independently of events leading to the exclusion of another region. The assumption was that despite events that would distort the interpretations of the eye data for a region (e.g., blinks, initial skipping), the region would be comprehended enough to be integrated in the rest of the sentence context and so not influence the processing of the other regions.

## 5.4 Results

### Reading task results

I first describe the results for the comprehension questions and the cognitive load task. These results indicate whether the participants carried out the reading task as instructed: reading for comprehension in the no-load condition and additionally retaining the load number in the load condition. For the latter, the combination of the comprehension question and load manipulation results also indicate whether the load task had the right level of difficulty (i.e., whether participants were able to fulfil both task satisfactorily at the same time).

After that, the trials included are listed for each region of interest, followed by the statistical analyses of the dependent measures for each region of interest.

### *Comprehension question and cognitive load results*

The number of correct and incorrect responses to the comprehension questions asking about parts of the sentences that were not gender-stereotype relevant in the no-load and load conditions as well as the entire sample can be found in Table 1.

Of the 2440 button press responses to the comprehension questions, 9 (0.37%) were excluded because one participant had not realised during her first block that she had to indicate responses with the button press device rather than just moving through the sentences.

Participants responded correctly on 2328 of the remaining 2431 trials (96% accuracy). The distribution of correct and incorrect responses was very similar for the no-load and load condition (see Table 1). The percentage of incorrect responses was for the no-load condition 4.6 and for the load condition 3.9. A one-sample *t*-test showed that the number of correct responses was significantly above chance ($t(60) = 88.96$, $p <$

.001). This result indicates that the participants had read the sentences in the reading task for comprehension, as instructed.

Table 1: Number of correct and incorrect responses to the comprehension questions per condition

| | No-load | Load | Total |
|---|---|---|---|
| Correct response | 1098 | 1230 | 2328 |
| Incorrect response | 53 | 50 | 103 |
| Total | 1151 | 1280 | 2431 |

Of the 1216 probe numbers, 925 (76%) were responded to correctly. A one-sample *t*-test revealed that the number of correct responses was significantly above chance ($t(31) = 15.36$, $p < .001$). This indicates that on most trials the participants kept the load numbers in mind during sentence reading to be able to compare them successfully to the probe numbers afterwards. The participants can therefore be assumed to have been under additional cognitive load as they read the sentences. This result combined with the high correct response rate to the comprehension questions suggests that the cognitive load was not too heavy, as participants were still able to comprehend the sentences.

### *Trials included in eye-data analysis*

Table 2 displays the number and percentages of trials included in the eye data analyses for each interest area region after the application of the exclusion criteria.

Table 2: Overview of number (and percentages) of trials included in the eye data analysis for the different interest areas

| | Agent region | Pronoun region | Spill-over region | Entire sentence region |
|---|---|---|---|---|
| All trials | 1464 (100%) | 1464 (100%) | 1464 (100%) | 1464 (100%) |
| After blink trials exclusion | 1450 (99.04%) | 1459 (99.66%) | 1457 (99.52%) | - |
| After skipped trials exclusion | 1330 (90.85%) | 1199 (81.90%) | 1096 (74.86%) | - |
| After outlier exclusion | 1327 (90.64%) | 1195 (81.63%) | 1091 (74.52%) | 1457 (99.5%) |

### *Statistical analysis*

Mixed-model analyses of variance (ANOVAs) by participants ($F_1$) and items ($F_2$)[12] were carried out for each interest region and each eye-movement measure with the within-participants factors match (match versus mismatch) and half (first half versus second half) and the between-participants factor load (no-load versus load). The factor half was included because it was possible that certain effects would decrease over the course of the experiment. An inspection of the data revealed that similar effects were found for both stereotypically female and male occupation labels. The factor gender was therefore not included in the analyses.

*The agent region*

Agent region means for the different eye-movement measures can be found in Tables 3 and 4. A table with the cell means and standard deviations used in the ANOVA participant analyses can be found in Appendix 5. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 6.

[12] One item had to be removed from all item analyses, because it erroneously appeared only in the first half of the experiment.

For the agent region, I expected no early effects because the match/mismatch had not yet occurred. I expected later effects if participants re-read the agent information more often in the mismatch than in the match condition.

As expected, no significant effect for match was found for the early viewing-time measures before the pronoun was encountered. Consistent with my expectations, the late measure Total Reading Time was significantly shorter in the match than in the mismatch condition ($F_1(1,59) = 8.30$, $p < .01$; $F_2(1,22) = 11.43$, $p = .01$). It was also significantly shorter in the second half than in the first half ($F_1(1,59) = 19.67$, $p < .001$; $F_2(1,22) = 27.45$, $p < .001$). The same pattern was found for the late measure Regression In: the proportion of regressions made into the region was smaller in the second half than in the first half ($F_1(1,59) = 6.67$, $p < .05$; $F_2(1,22) = 5.83$, $p < .05$). The effects for match show that participants looked back to the agent significantly more often after they had encountered a mismatching than a matching pronoun. The main effect of half on the late measures suggests that participants realised that they would repeatedly come across stereotype-mismatching information and felt less need to reconfirm the mismatch.

There was also a significant interaction between match and half for the Regression In measure ($F_1(1,59) = 5.33$, $p < .05$; $F_2(1,22) = 4.27$, $p = .051$) with a greater difference between the match and mismatch conditions in the first half compared to the second half. Planned comparisons showed that the effect for match was only significant in the first half ($F_1(1,59) = 6.21$, $p < .05$; $F_2(1,22) = 5.42$, $p = .05$). This confirms that participants did look back into the region more often after encountering a mismatching than a matching pronoun, but did this less over time and after several mismatching encounters.

No main effect of load was found on any of the dependent variables. There were, however, interactions between cognitive load and match on the early measures First Fixation Duration and First-Pass Duration. In the no-load condition, processing times were longer in the match than in the mismatch condition, whereas in the load condition, processing times were shorter in the match than in the mismatch condition. Posttests, however, showed that the difference was only significant in the load condition. No account of this pattern could be offered given that the pronoun conveying the matching or mismatching information had not even been encountered.

Table 3: Eye-movement measure means for the agent region by load condition[13]

| | No-load | | | Load | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-match | Total | Match | Mis-match | Total |
| First Fix. Duration | 218 | 206 | 212 | 207 | 218 | 213 | 213 | 212 | 212 |
| First-Pass Duration | 287 | 275 | 281 | 263 | 281 | 272 | 275 | 278 | 277 |
| Selective Reg.-Path Duration | 312 | 316 | 314 | 293 | 299 | 296 | 302 | 307 | 305 |
| Total Reading Time | 389 | 417 | 403 | 358 | 407 | 383 | 373 | 412 | 393 |
| Regression Out | 0.15 | 0.18 | 0.16 | 0.19 | 0.14 | 0.16 | 0.17 | 0.16 | 0.16 |
| Regression In | 0.19 | 0.22 | 0.21 | 0.20 | 0.23 | 0.21 | 0.20 | 0.23 | 0.21 |

[13] In all tables: First Fixation Duration, First-Pass Duration, Selective Regression-Path Duration, and Total Reading Time are indicated in milliseconds. Regression Out and Regression In are indicated in proportion of valid trials.

Table 4: Eye-movement measure means for the agent region by halves

| | Half 1 | | | Half 2 | | |
|---|---|---|---|---|---|---|
| | Match | Mis-Match | Total | Match | Mis-Match | Total |
| First Fix. Duration | 211 | 215 | 213 | 214 | 210 | 212 |
| First-Pass Duration | 283 | 285 | 284 | 267 | 271 | 269 |
| Selective Reg.-Path Duration | 311 | 317 | 314 | 293 | 298 | 296 |
| Total Reading Time | 396 | 458 | 427 | 351 | 366 | 358 |
| Regression Out | 0.18 | 0.16 | 0.17 | 0.15 | 0.16 | 0.15 |
| Regression In | 0.20 | 0.28 | 0.24 | 0.19 | 0.17 | 0.18 |

*The pronoun region*

Pronoun region means for the different eye-movement measures can be found in Tables 5 and 6. A table with the cell means and standard deviations used in the ANOVA can be found in Appendix 7. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 8.

For the pronoun region, match effects were expected during word recognition and early processing as well as during later processes of contextual integration.

The data confirmed these expectations with significantly shorter viewing times for both early and late eye-movement measures in the match than in the mismatch condition: First Fixation Durations ($F_1$(1,59) = 9.81, $p$ < .005; $F_2$(1,22) = 9.11, $p$ < .01), First-Pass Durations ($F_1$(1,59) = 7.95, $p$ < .01; $F_2$(1,22) = 11.72, $p$ < .005), Selective Regression-Path Durations ($F_1$(1,59) = 5.62, $p$ < .05; $F_2$(1,22) = 10.87, $p$ <

.005) and Total Reading Time ($F_1$(1,59) = 22.02, $p$ < .001; $F_2$(1,22) = 30.27, $p$ < .001).

For the proportion of first-pass regressions out of the pronoun region (early measure), no significant differences were found between the match and mismatch condition. The number of regressions into the region was significantly greater in the mismatch than in the match condition ($F_1$(1,59) = 6.62, $p$ < .05; $F_2$(1,22) = 8.48, $p$ < .05).

Total Reading Time and Regression In are considered late measures, capturing processes that carry on after a region has been left. The effect of match suggests that the processing of the mismatching information carried on more than the processing of the matching information.

No differences were found between the no-load and load conditions, suggesting that cognitive load had no influence on early mismatch recognition or later mismatch resolution.

Table 5: Eye-movement measure means for the pronoun region by load condition

| | No-load | | | Load | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-match | Total | Match | Mis-match | Total |
| First Fix. Duration | 211 | 225 | 218 | 207 | 223 | 215 | 209 | 224 | 216 |
| First-Pass Duration | 236 | 258 | 247 | 222 | 239 | 231 | 229 | 248 | 239 |
| Selective Reg.-Path Duration | 251 | 273 | 262 | 236 | 252 | 244 | 243 | 262 | 253 |
| Total Reading Time | 314 | 393 | 354 | 312 | 371 | 341 | 313 | 382 | 348 |
| Regression Out | 0.10 | 0.10 | 0.10 | 0.10 | 0.12 | 0.11 | 0.10 | 0.11 | 0.11 |
| Regression In | 0.18 | 0.27 | 0.23 | 0.25 | 0.30 | 0.27 | 0.22 | 0.28 | 0.25 |

Table 6: Eye-movement measure means for the pronoun region by halves

| | Half 1 | | | Half 2 | | |
|---|---|---|---|---|---|---|
| | Match | Mis-Match | Total | Match | Mis-match | Total |
| First Fix. Duration | 206 | 224 | 215 | 212 | 223 | 217 |
| First-Pass Duration | 229 | 247 | 238 | 229 | 250 | 240 |
| Selective Reg.-Path Duration | 245 | 258 | 252 | 242 | 267 | 254 |
| Total Reading Time | 318 | 402 | 360 | 308 | 362 | 335 |
| Regression Out | 0.12 | 0.10 | 0.11 | 0.08 | 0.12 | 0.10 |
| Regression In | 0.22 | 0.33 | 0.27 | 0.22 | 0.24 | 0.23 |

*The pronoun spill-over region*

Pronoun spill-over region means for the different eye-movement measures are shown in Tables 7 and 8. A table with the cell means and standard deviations used in the ANOVA[14] can be found in Appendix 9. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 10.

For the pronoun spill-over region, match effects were expected to arise if processing of the pronoun continued during reading the next word or group of words. Such a spill-over effect was found on the Selective Regression-Path Duration, with longer viewing times in the mismatch than in the match condition. This effect was marginal by participants ($F_1(1,59) = 3.45$, $p = .067$), and significant by items ($F_2(1,21) = 5.29$, $p < .05$). The proportion of first-pass regressions out of the pronoun spill-over region

[14] Only 22 items were included in the item analysis, as the value was missing for one item in one condition due to exclusion of trials.

(early measure) was significantly greater in the mismatch than in the match condition: ($F_1(1,59) = 10.52$, $p < .005$; $F_2(1,21) = 10.53$, $p < .01$), indicating that the number of times participants made at least one fixation out of the spill-over region after first encountering it was greater when they had just read a mismatching rather than a matching pronoun. This suggests that participants still needed additional processing time when they first left a mismatching pronoun region. This is supported by the fact that the proportion of regressions into the pronoun region was greater in the mismatch than the match condition.

There were also effects of half on the late measures Total Reading Time and Regression In. The Total Reading Time was greater for the first half than for the second half ($F_1(1,59) = 4.19$, $p < .05$; $F_2(1,21) = 4.31$, $p = .05$). In addition, significantly more regressions into the region were made in the first than in the second half ($F_1(1,59) = 6.78$, $p < .05$; $F_2(1,21) = 4.66$, $p < .05$). These late effects might have emerged because the pronoun spill-over region is quite close to the end of the sentence. At this point, participants might have wanted to reconfirm the contextual information of the end region, before moving on to the comprehension question. They might have felt the need to do so more in the first half of the experiment, when they were still getting accustomed to the task, than in the second half.

No effects of load were found on the viewing-time measures. However, the proportion of first-pass regressions (early measure) made out of the pronoun spill-over region was significantly greater in the load than the no-load condition ($F_1(1,59) = 5.96$, $p < .05$; $F_2(1,21) = 10.94$, $p < .05$). These backwards regressions only took place after the most important information had been read (agent, reflexive verb, and reflexive pronoun). Participants looked back to reread this information more in the load than in the no-load condition.

Table 7: Eye-movement measure means for the pronoun spill-over region by load condition

| | No-load | | | Load | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-match | Total | Match | Mis-match | Total |
| First Fix. Duration | 231 | 234 | 233 | 233 | 226 | 230 | 232 | 230 | 231 |
| First-Pass Duration | 284 | 288 | 286 | 270 | 286 | 278 | 277 | 287 | 282 |
| Selective Reg.-Path Duration | 318 | 324 | 321 | 299 | 335 | 317 | 309 | 329 | 319 |
| Total Reading Time | 405 | 412 | 408 | 378 | 407 | 393 | 391 | 409 | 400 |
| Regression Out | 0.14 | 0.21 | 0.18 | 0.21 | 0.33 | 0.27 | 0.18 | 0.27 | 0.22 |
| Regression In | 0.23 | 0.19 | 0.21 | 0.20 | 0.19 | 0.20 | 0.22 | 0.19 | 0.20 |

Table 8: Eye-movement measure means for the pronoun spill-over region by halves

| | Half 1 | | | Half 2 | | |
|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-match | Total |
| First Fix. Duration | 229 | 231 | 230 | 236 | 229 | 232 |
| First-Pass Duration | 271 | 295 | 283 | 283 | 279 | 281 |
| Selective Reg.-Path Duration | 295 | 342 | 318 | 323 | 317 | 320 |
| Total Reading Time | 399 | 437 | 418 | 384 | 382 | 383 |
| Regression Out | 0.16 | 0.29 | 0.22 | 0.20 | 0.25 | 0.23 |
| Regression In | 0.25 | 0.22 | 0.24 | 0.18 | 0.16 | 0.17 |

*The entire sentence region*

Entire sentence region means for the different eye-movement measures can be found in Tables 9 and 10. A table with the cell means and standard deviations used in the ANOVA can be found in Appendix 11. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 12.

I expected sentence viewing times to be longer in the mismatch than in the match condition, reflecting late comprehension and context integration processes. Indeed, processing time assessed by the late measure Sentence Reading Time was significantly greater in the mismatch than in the match condition ($F_1(1,59) = 22.43$, $p < .001$; $F_2(1,22) = 25.64$, $p < .001$). This suggests that the stereotype-mismatching information influenced the comprehension process not only locally, where it occurred, but also during the later processing on the sentence level.

The Total Reading Time was significantly shorter in the second half than in the first half ($F_1(1,59) = 63.04$, $p < .001$; $F_2(1,22) = 60.03$, $p < .001$), which indicates that participants became accustomed to the task and the type of sentences they would encounter throughout the experiment and therefore took less time to read and process the information. This interpretation of the Sentence Reading Time results, considering the main effects of match and half, is supported by the results of the other late sentence comprehension measure Sentence Fixation Count: Significantly more fixations were made in the mismatch than in the match condition ($F_1(1,59) = 19.12$, $p < .001$; $F_2(1,22) = 16.66$, $p < .001$) and significantly more fixations were made in the in the first half than in the second half ($F_1(1,59) = 52.83$, $p < .001$; $F_2(1,22) = 37.98$, $p < .001$).

Cognitive load had a significant effect on the Sentence Reading Time, with *shorter* viewing times in the load than in the no-load condition ($F_1(1,59) = 5.02$, $p < .05$;

$F_2$(1,22) = 73.08, $p < .001$). Also, significantly fewer fixations were made in the load compared to the no-load condition (13.96; $F_1$(1,59) = 7.36, $p < .01$; $F_2$(1,22) = 81.72, $p < .001$).

This effect of cognitive load on both measures on the entire sentence level indicates that participants used different reading strategies for the no-load and load condition. Because they had to retain a 5-digit number in the load condition, they seem to have tried to complete the reading task quickly in order to minimise the time they had to maintain the number. This assumption is supported by the comments participants made after the experiment about their number retention strategy.

Table 9: Eye-movement measure means for the entire sentence region by load condition

| | No-load | | | Load | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-match | Total | Match | Mis-match | Total |
| Sentence Reading Time | 3011 | 3173 | 3092 | 2583 | 2767 | 2675 | 2797 | 2970 | 2883 |
| Sentence Fixation Count | 13.61 | 14.31 | 13.96 | 11.52 | 12.29 | 11.91 | 12.56 | 13.30 | 12.93 |

Table 10: Eye-movement measure means for the entire sentence region by halves

| | Half 1 | | | Half 2 | | |
|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-Match | Total |
| Sentence Reading Time | 2966 | 3202 | 3084 | 2627 | 2738 | 2683 |
| Sentence Fixation Count | 13.21 | 14.16 | 13.69 | 11.92 | 12.44 | 12.18 |

*Summary of eye-movement results*

An overview of the theoretically most interesting significant main effects and interactions for viewing time and proportion regression measures for the agent, pronoun, and pronoun spill-over regions can be found in Table 11. An overview of the theoretically most interesting significant main effects and interactions for the eye-movement measures for the entire sentence region can be found in Table 12. *X* marks a significant main effect with *p* < .05 by participants and items.

Table 11: Summary of eye-movement results for the agent, pronoun and pronoun spill-over regions for the factors match and load[15]

| | Agent region | | | Pronoun region | | | Pronoun spill-over region | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Load | Match x Load | Match | Load | Match x Load | Match | Load | Match x Load |
| First Fix. Duration | - | - | X | X | - | - | - | - | - |
| First-Pass Duration | - | - | X | X | - | - | - | - | - |
| Selective Reg.-Path Duration | - | - | - | X | - | - | X[16] | - | - |
| Total Reading Time | X | - | - | X | - | - | - | - | - |
| Regression Out | - | - | - | - | - | - | X | X | - |
| Regression In | - | - | - | X | - | - | - | - | - |

[15] For all summary tables, "x" marks a significant effect (*p* < .05) by participants and items unless otherwise indicated.

[16] Marginal by participants (p = .067)

Table 12: Summary of eye-movement results for the entire sentence region for the factors match and load

| | Match | Load | Match x Load |
|---|---|---|---|
| Sentence Reading Time | X | X | - |
| Sentence Fixation Count | X | X | - |

## Memory tasks results

### *Gender cued-recall questionnaire*

The number of matching and mismatching responses in response to matching or mismatching stimuli in the reading task for both the no-load and load condition can be found in Table 13. Three values were missing from the analyses, one in the no-load condition and two in the load condition). Correct responses are marked with (c), incorrect responses with (i).

For the analyses of the gender cued-recall questionnaire, I applied Signal Detection Theory to determine the statistics of memory sensitivity (*d'*)[17] and response bias (*C*)[18] separately (see Macmillan & Creelman, 1991). Memory sensitivity measures how many of the participants' responses corresponded to the presented information, which assesses how much the participants remembered. An estimate can be derived by comparing the positive diagonals (number of correct responses) versus the negative diagonals (number of incorrect responses) in Table 13. Bias measures the strength of the participants' tendency to respond in a stereotype-matching or -mismatching way, independently of which information had been presented. An estimate can be derived

[17] *d'* for discrimination: the difference between the means of the distributions of signal present and signal absent

[18] *C* for decision criterion

by comparing the columns for matching and mismatching responses in Table 13. The calculations of *C* and *d’* use the hit and false-alarm rate of the response data. The hit-rate is the rate of matching responses made after the presentation of a matching stimulus (e.g., the first cell in the first row in Table 13), and the false-alarm rate is the rate of matching responses made after the presentation of a mismatching stimulus (e.g., first cell in the second row of Table 13).

*d*' is calculated by subtracting the *z*-score of the false-alarm rate from the *z*-score from the hit rate. A *d’* value of 0 indicates no memory sensitivity, and greater positive values indicate greater memory sensitivity (with *d’* = 4.65 widely considered as the effective ceiling value). *C* is calculated by multiplying the sum of the z-scores of the false-alarm rate and the z-scores of the hit rate with (-0.5). A *C* value of 0 indicates no response bias, negative values indicate an expectancy consistency (conservative) bias, and positive values indicate an expectancy inconsistency (liberal) bias; extreme values of *C* are +/-2.33 (see Macmillan & Creelman, 1991).

Memory sensitivity was found to be $d'_{load}$ = 1.12 in the load condition and $d'_{no\text{-}load}$ = 0.91 in the no-load condition. Response bias was found to be $C_{load}$ = -0.53 in the load, and $C_{no\text{-}load}$ = -0.30 in the no-load condition, indicating a conservative load bias in both conditions.

Table 13: Number of matching and mismatching responses in response to matching or mismatching items in the gender cued-recall questionnaire[19]

| | No-load | | Load | | Total | |
|---|---|---|---|---|---|---|
| | Matching responses | Mismatch. responses | Matching responses | Mismatch. responses | Matching responses | Mismatch. responses |
| Matching item | 279 (c) | 68 (i) | 320 (c) | 62 (i) | 599 (c) | 130 (i) |
| Mismatch. item | 138 (i) | 210 (c) | 203 (i) | 181 (c) | 341 (i) | 391 (c) |
| SUM | 417 (60%) | 278 (40%) | 523 (68%) | 243 (32%) | 940 (64%) | 521 (36%) |

In order to determine whether the sensitivity and bias were significantly different from 0 and whether the differences between the no-load and load conditions were significant, I carried out a log-linear analysis[20].

A hierarchical fully saturated log-linear analysis was applied to the present data set with the factors item (match, mismatch), response (match, mismatch) and load (no-load, load). The interactions included the two-way interactions Item x Response, Item x Load and Response x Load and the three-way interaction Item x Response x Load. The two-way interaction Item x Response is another formulation for memory sensitivity; the factor response is another formulation for response bias. That means that if the interaction between load and Item x Response is a necessary factor for the appropriate model of the data, cognitive load has a significant effect on memory

[19] In all tables: (c) indicates a correct response; (i) indicates an incorrect response

[20] Log-linear analysis is a goodness-of-fit test which determines the model that best represents a set of data. It starts off comparing the data with the saturated model, which includes the main effects of all factors and all levels of interactions (Howitt & Cramer, 2005). A backward elimination procedure removes step by step any factors or interactions which do not increase the goodness-of-fit of the model to the data, starting with the lowest level interactions. The factors or interactions are indicated by non-significant p-values. The model that best fits the data contains only factors or interactions that would decrease the fit of the model when removed. They are indicated by significant p-values.

sensitivity. Similarly, if the interaction between load and response is a necessary factor for the appropriate model of the data, cognitive load has a significant effect on response bias.

The log-linear analyses revealed that an appropriate model of the data does not require the three-way Item x Response x Load interaction ($G^2(1) = 1.55$; $p = .21$) or the two-way Item x Load interaction ($G^2(1) = 1.82$; $p = .18$). It does, however, require the Item x Response two-way interaction ($G^2(1) = 208.58$; $p < .001$) and Response x Load condition interaction ($G^2(1) = 10.88$; $p = .001$). These interactions are included in the final formulation of the model along with the factors that yielded main effects: item, response and load.

The result that the interaction Item x Response is a required factor of the model shows that the participants' responses depended on which stimuli they had seen (compare 990 total number of correct responses with 471 total number of incorrect responses in Table 13). As the interaction with load was not required for a model of the data, memory sensitivity did not differ significantly between the load conditions.

The result of response being a required factor of the model shows that there was a significant response bias. The descriptive data showed that it was conservative towards expectancy consistency in both load conditions. The result that the Response x Load interaction is required for a model of the data shows that the bias was affected by load. As has been seen in the descriptive data, the bias was greater in the load than in the no-load condition ($C_{load} = -0.53$ versus $C_{no\text{-}load} = -0.30$). That means that, independently of the items presented, the percentage of matching responses in comparison to mismatching responses was significantly greater in the load condition (68% versus 32%) than in the no-load condition (60% versus 40%, see Table 13).

Overall, the results of the gender-cued recall questionnaire show that participants were able to identify the gender of the agents they had read about well above chance level. This ability was not affected by cognitive load in the encoding phase. Furthermore, participants displayed a conservative response bias towards stereotype-matching responses and did so even more when they were under additional cognitive load during the encoding phase. These results will be taken up again in the discussion.

### *Sentence-memory questionnaire*

The number of correct and incorrect responses for matching and mismatching items in the sentence-memory questionnaire, administered in the load condition only, can be found in Table 14.

Participants identified the sentence endings significantly more often correctly than incorrectly as shown by a one-sample *t*-test: $t(31) = 37.80$; $p < .001$. Thus, participants had good overall memory of the information presented in the sentences, despite the cognitive load during encoding. Sentence memory did not differ for matching or mismatching sentences.

Table 14: Number of correct and incorrect responses for matching or mismatching items in the sentence-memory questionnaire (load condition only)

| | Correct responses | Incorrect responses |
|---|---|---|
| Matching items | 360 | 24 |
| Mismatching items | 347 | 37 |
| SUM | 707 | 61 |

## 5.5 Discussion

The first goal of this study was to replicate the stereotype-mismatch effect found in previous studies (Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Sturt, 2003) as a basis for its further exploration. The second goal was to investigate whether the recognition and resolution of the agent-pronoun mismatch was capacity demanding or whether it still took place under cognitive load. The third goal was to examine the participants' memory for stereotype-relevant information as an indication of the representation constructed during online processing depending on the cognitive capacities during encoding.

**Evidence for stereotype-mismatch effects on online measures**

Both early and late eye-movement measures showed that the mismatch effect was indeed replicated. For the agent region, there was a mismatch effect for the late viewing-time measure Total Reading Time. For the pronoun region, there was a mismatch effect on all viewing time and the Regression In measures. For the spill-over region there was a mismatch effect for the Regression Out measure and for the late viewing-time measure Selective Regression-Path Duration (significant by items and marginally significant by participants).

This result pattern suggests the following chronological order of processing events, relevant to the stereotypical information. Participants needed more processing time when first encountering a stereotype-mismatching pronoun than when encountering a matching pronoun. They also looked back more to the agent region after reading a mismatching than a matching pronoun, as if to reconfirm the information they had read. Effects on later measures on the pronoun region show that participants allocated more processing time to mismatching pronouns later during sentence processing. The spill-over effect on the Regression Out measure suggests that participants looked back into earlier sentence regions, most likely the pronoun region, when they needed more processing time after encountering a mismatching than a matching pronoun.

On the sentence level, as expected, the reading time was significantly longer in the mismatch than in the match condition (2970 msec versus 2797 msec). This difference of 173 msec is considerably greater than the difference of 69 msec for the Total Reading Time for the pronoun region. This suggests that the stereotype-mismatch influenced the comprehension process not only on a local level, where the mismatch occurred, but also on a more global sentence level where late context integration processes take place. This fits with the claim that "the effects of higher-order language processing are often delayed and/or apparent over a wider temporal window than are the effects of lower-order language processing" (Reichle, Rayner, & Pollatsek, 2003).

A review of the present and previous findings (see Table 15) shows that there are a few differences in when the effect emerged. In the present experiment, as in Sturt (2003), the effect on the pronoun emerged early, on the First Fixation Durations, as well as on later measures. An effect on the spill-over region emerged in Sturt's study on the late measure Regression-Path Duration[21], but not the earlier measures. In the first experiment by Duffy and Keir (2004), the difference between matching and mismatching pronouns was significant only by items in the later measure Regression-Path Duration, and spilt over to the post-pronoun region, gaining significance by items only in the First-Pass Duration and by items and participants in the Regression-

[21] Regression-Path Duration includes all fixations and refixations made on a region and during regressions to the left until the eyes leave the region to the right.

Path Duration and Second Pass Time[22]. Kreiner, Sturt and Garrod (2008), too, found the earliest mismatch effect on the pronoun in the Regression-Path Duration, with the effect spilling over to early as well as late measures in the spill-over region. Taken together, these results indicate the importance of including the spill-over region in the analyses. Apparently, there can be a trade-off between the main region of interest and the spill-over region. In studies where early effects have been found on the pronoun region, no early effects have been observed on the viewing times on the spill-over region. In one of the experiments where no early effects arose on the pronoun region (Kreiner, Sturt, & Garrod, 2008), there were early effects on the spill-over region. In another experiment without early effects on the pronoun region (Duffy & Keir, 2004), however, there were no such early effects on the spill-over region.

Reichle, Pollatsek, Fisher and Rayner (1998) define spill-over[23] as “an effect of processing a given word that occurs after fixating that word” (p. 145). In their E-Z Reader model the initiation of an eye-movement is dependent on “the successful completion of a psychological process (such as lexical access)” (p. 129). Reichle and colleagues divide lexical access into the familiarity stage and the stage of completing lexical access and model the mean durations of both stages to be dependent on factors like the word’s frequency. According to this model, spill-over effects arise when the duration of the stage of completing lexical access to word *n* is increased and therefore the preview on the word *n+1* while fixating word *n* is decreased. This results in increased viewing times on the word *n+1* when it is fixated. It is not obvious why the spill-over effects differ between the studies described above, given that the same

[22] “Second pass time is the time spent refixating a region after the eyes have left the region.” (Duffy and Keir, 2004, p. 554)

[23] Spill-over in the E-Z model (Reichle, Pollatsek, Fisher, & Rayner, 1998) only refers to effects on early viewing-time measures. Effects on later measures are assumed to reflect higher level comprehension and wrap-up processes at the end of the sentence.

words were used for the pronoun region, keeping lexical variables constant. They may be due to higher level processes not captured in most versions of the E-Z reader model (but see Reichle, Warren, & McConnell, 2009).

An eye-movement measure that has not been compared here for the stereotype-matching and mismatching sentences is the probability of word skipping. As it has sometimes been used in the past though to reflect processing difficulty along with viewing-time measures (e.g., White, Rayner, & Liversedge, 2005), I will briefly discuss the possible word skipping effects in this Experiment.

Previous studies have identified some variables influencing the likelihood of a word being skipped. According to these findings, short words are more likely to be skipped than long words, frequent words are more likely to be skipped than infrequent words and words that are more predictable from the preceding context are more likely to be skipped than words that are less predictable from the preceding context (for a review see Brysbaert, Drieghe, & Vitu, 2005). In the present experiment, the only such variable that differed between the stereotype-matching and –mismatching sentences was the predictability of the pronoun. In the sentence "Last week the secretary familiarized *herself* with the new photocopier" the pronoun was, for example, more predictable from the context than in the sentence "Last week the secretary familiarized *himself* with the new photocopier". Drieghe, Rayner and Pollatsek (2005) found that a highly predictable word like "liver" in the sentence "The doctor told Fred that his drinking would damage his *liver* very quickly" was skipped more often than a less predictable word like "heart" in the sentence "The doctor told Fred that his drinking would damage his *heart* very quickly". Based on this finding, it could be expected that in Experiment 1, more predictable stereotype-matching pronouns would be more often skipped than less predictable stereotype-mismatching pronoun. This finding would be in accordance with my eye-movement findings.

Table 15: Overview of significant (p<.05) mismatch effect findings on the pronoun and spill-over region in the present and previous studies

| Study | Example | Effects on the pronoun region | Effects on the spill-over region |
|---|---|---|---|
| Present study, Experiment 1 | "Last week the secretary familiarised herself/himself with the new photocopier." | First Fixation Duration,<br>First-Pass Duration,<br>Selective Regression-Path Duration,<br>Total Reading Time | Selective Regression-Path Duration |
| Sturt (2003), Experiment 1 | " [Prior context]<br>He remembered that the surgeon had pricked himself/herself with a used syringe needle. [Subsequent context] " | First Fixation Duration,<br>First-Pass Duration,<br>Regression-Path Duration,<br>Second Pass Time | Regression-Path Duration |
| Duffy and Keir (2004),<br>Experiment 1 | "The babysitter found herself/himself humming while walking up the door." | - | Regression-Path Duration,<br>Second Pass Time |
| Duffy and Keir (2004),<br>Experiment 2 | " [Prior neutral context]<br>The electrician taught himself/herself a lot while fixing the problem. " | First-Pass Duration,<br>Regression-Path Duration | Regression-Path Duration |
| Kreiner, Sturt and Garrod (2008), Experiment 1 | " [Title]<br>Yesterday the minister left London after reminding himself/herself about the letter." | Regression-Path Duration,<br>Second Pass Time | First Fixation Duration,<br>First-Pass Duration,<br>Regression-Path Duration |

### Effects of cognitive load on mismatch detection and resolution

My expectation was that cognitive load should slow down the overall reading process. I found, however, that the overall sentence reading time was faster and fewer fixations were made in the load than in the no-load condition. The reason might be that participants had to keep the load number in mind while reading and they might have tried to complete the reading task quickly to minimise the time they had to retain the number.

Although cognitive load affected the total reading time, it did not have a main effect on the viewing-time measures for the agent or the pronoun, apart from an effect for Regression Out of the spill-over region with more regressions in the load than the no-load condition. An explanation for this could be that cognitive load only affected late stages of the comprehension process. The spill-over region marks the end of the most informative regions of a sentence, including the agent region, the reflexive verb and the reflexive pronoun region. Cognitive load did not affect the local processing of these regions but maybe it did affect the more strategic processes of integrating the crucial sentence information at a later stage of the sentence assembly process.

The absence of systematic effects of working memory load on early measures of reading time suggests the processes captured by these measures (e.g., early word recognition, lexical access) are relatively automatic.

The idea that early processes of the comprehension process take place relatively automatically, whereas the later, more capacity-demanding processes seem more likely to be affected by strategic processes, is supported by the effects of half. No early effects of half were observed, but there were some later main effects (agent region: Total Reading Time, Regression In; spill-over region: Total Reading Time; Regression In; entire sentence region: Total Reading Time, Total Fixation Count) and an interaction with match (agent region: Regression In). The late measures might reflect strategic processes taking place more in the first than the second half of the experiment, for example, looking back to reconfirm the agent’s occupation after encountering a mismatching pronoun.

The focus of the present study was not on main effects of cognitive load on the reading process but on their potential interactions with effects of match. From previous psycholinguistic studies using a similar reading task (Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Sturt, 2003), it had not been clear whether the recognition and resolution of the agent-pronoun mismatch is capacity-demanding or whether it still takes place under cognitive load. I had expected that the readers might be less likely to detect or resolve the mismatch under cognitive load. However, no interaction was observed for any of the eye-movement measures. This finding is, however, difficult to interpret as overall, cognitive load did not have the expected effect of slowing down reading. Nevertheless, it can be concluded that readers probably read quickly to unload the memory load, but this did not affect their ability to detect and solve the mismatch.

### Memory for stereotype-relevant information

The analyses of the gender questionnaire showed that the cued-recall of the stereotype-relevant information was significantly above chance regardless of cognitive load. These results indicate that participants had constructed correct representations during reading. This finding contributes to an explanation of the reading data differences between matching and mismatching sentences: Upon encountering a mismatching pronoun, participants spent extra time not only detecting the mismatch but also resolving it.

The analyses also revealed that the participants tended to respond in a stereotype-matching way when asked for the agents' gender. This conservative response bias increased under cognitive load. Apparently, the participants employed guessing strategies when they could not remember an agent's gender, taking into consideration the information the gender stereotype provides. They did this even more when they were under additional cognitive load.

This finding suggests that the load manipulation affected one aspect of cued-recall memory: the bias. Unfortunately, the load variable was confounded with a difference in the duration of the experimental sessions in the no-load and load condition (30 versus 45 minutes)[24]. The greater response bias in the load condition might therefore originate from the longer time interval between encoding and recalling the information.

However, if the difference in the length of the session caused the effect on the bias, there should be less bias for the second than for the first half. I calculated the response bias across load conditions for each experimental half separately and found it to be $C_{half1}$ = -0.40 in the first half and $C_{half2}$ = -0.43 in the second half. The memory sensitivity was $d'_{half1}$ = 1.12 in the first half and $d'_{half2}$ = 0.9 in the second half. The number and percentage of matching and mismatching responses in response to matching or mismatching items divided by halves are displayed in Table 16.

A log-linear analysis with the factors item (match, mismatch), response (match, mismatch), and half (half1, half2) revealed that an appropriate model of the data does not require the three-way Item x Response x Half interaction ($G^2(1) = 2.17$; $p = .14$), the two-way Item x Half interaction ($G^2(1) = 0.023$; $p = .88$), the two-way Response x Half interaction ($G^2(1) = 5.38$; $p = .46$) or the main effect of half ($G^2(1) = 0.001$; $p = .98$). It only requires the Item x Response interaction ($G^2(1) = 208.58$; $p < .001$) and the factors that yielded a main effect: item and response. This shows that the results in

[24] Initially the primary goal of this study was the online processing effects. Otherwise greater care would have been taken to control the duration of the experiment.

the first and second halves did not differ from each other and that, in particular, the interaction between response and half was not significant (response is another formulation for bias; see section 5.4). This finding argues against the account that the bias was greater in the load condition because of the longer experimental sessions and makes it more likely that it occurred because of the cognitive load manipulation during encoding.

From previous studies, it had not been clear which processes were reflected in the increased processing time to the stereotype-mismatching over -matching information. The memory findings in this study demonstrate that readers resolve the mismatches sufficiently during online processing to have above chance accurate memory later. However, memory was not perfect. This could either indicate that not all mismatches had been resolved (enough) or that the information had been forgotten by the time participants filled in the memory questionnaire. This question was followed up in Chapter 3.

Table 16: Number and percentage of matching and mismatching responses in response to matching or mismatching items in the gender cued-recall questionnaire divided by halves

| | No-load | | | Load | | |
|---|---|---|---|---|---|---|
| | Matching responses | Mismatching responses | Total | Matching responses | Mismatching Responses | Total |
| Matching item half 1 | 144 (c)<br>84% | 28 (i)<br>16% | 172 | 158 (c)<br>83% | 33 (i)<br>17% | 191 |
| Mismatching item half 1 | 61 (i)<br>35% | 114 (c)<br>65% | 175 | 100 (i)<br>52% | 92 (c)<br>48% | 192 |
| Total half 1 | 205<br>59% | 142<br>41% | 347 | 258<br>67% | 125<br>33% | 383 |
| Matching item half 2 | 135 (c)<br>77% | 40 (i)<br>23% | 175 | 162(c)<br>85% | 29 (i)<br>15% | 191 |
| Mismatching item half 2 | 77 (i)<br>45% | 96 (c)<br>55% | 173 | 103 (i)<br>54% | 89 (c)<br>46% | 192 |
| Total half 2 | 212<br>61% | 136<br>39% | 348 | 265<br>69% | 118<br>31% | 383 |
| Total | 417<br>60% | 278<br>40% | 695 | 523<br>68% | 243<br>32% | 766 |

### Interpretation of the findings within the working model

As described in section 3, my working model consists of episodic exemplar and semantic prototype representations, organised in an associative network. When a new sentence is read, an exemplar representation of the agent is constructed by connecting different nodes. If the gender of the agent matches the prototypical representation, this is a straight-forward process: When the agent region is first encountered, the occupation label (e.g., *secretary*) activates connected nodes or features - for example, the prototypical gender (e.g., *female*). This results in an expectation of the secretary to be female and in the formation of a temporary exemplar representation, connecting a *secretary* node to a *female* node. If this expectation is confirmed in the rest of the sentence, particularly after the pronoun has been encountered, then a more stable exemplar representation is established. However, if the gender of the agent, mismatches the prototypical representation, then forming an exemplar representation is more complicated. If the gender in the temporary exemplar representation is disconfirmed by the pronoun, then the exemplar representation has to be revised and a new connection has to be constructed between, for example, a *secretary* node and a *male* node. Assuming that detecting the expectancy mismatch and revising the representation is time-consuming, the mismatch effect is explained.

My results suggest that the recognition of a prototype match or mismatch and the formation of a prototype-matching or -mismatching exemplar representation can still take place when cognitive resources are limited, as reflected in the absence of an interaction between the mismatch effect and cognitive load.

Generally, when the episodic exemplar representation of a particular agent — for example, a particular secretary — can later not be remembered, the prototype representation for secretary is used as a reconstruction aid. If the gender of the exemplar and the prototype representations match, this results in an accurate memory performance. If, however, the gender of the two representations mismatch, memory performance will be inaccurate. This interpretation can explain the empirical finding of a conservative response bias: When participants could not remember a particular agent's gender, they seem to have used the stereotypical prototype representation as guessing aid. The finding that the bias was stronger in the load condition can be interpreted as result of the processing by-product principle (Carlston & Smith, 1996 as cited in Smith, 1998): It is assumed that the ease of reconstructing an exemplar representation corresponds to the effort allocated during its formation. Because this effort is limited by the cognitive resources available, participants had increased difficulties reconstructing the accurate exemplar representations under cognitive load conditions.

## Conclusions

The results of this study make a number of contributions to psycholinguistic theory. First, they confirm that the stereotype-mismatch effect found in previous studies is stable and replicable. Second, they show that working memory load leaves the reading process relatively unaffected (see also Chapter 5), but can influence memory performance for stereotype-relevant information later on.

The results also add to the social psychological literature. To date, studies investigating the effect of cognitive load on the processing of and memory for stereotype-relevant information have generally used impression formation tasks (e.g., Sherman, Lee, Bessenoff & Frost, 1998; Sherman & Frost, 2000). The information processing in the present study, however, resembles more casual reading than goal-directed impression formation. It has been shown here that accurate stereotype-mismatching representations can be constructed simply by reading about a stereotype-mismatching agent. During casual reading attention is often divided, as was the case with the cognitive load manipulation of the present study. When reading the newspaper, for example, people often listen to the radio at the same time. The present findings suggest that readers are still likely to pick up on stereotype-mismatching information.

The memory results in this study are thus particularly informative about the processes of updating or changing the representations of stereotyped groups. In my working model, the prototype representation is an abstraction of the relevant exemplar representations. Therefore, a change to the stereotype-consistent prototype can happen when a sufficient number of stereotype-mismatching exemplar representations have been formed. It is therefore important to see that the mere reading of a piece of mismatching information about a stereotype-mismatching agent can result in the formation of an episodic memory that can be recalled later at above chance level. If, however, the episodic representation cannot be reconstructed, memory will be biased towards the stereotypically expected way. This bias might be a functional mechanism of the information processing system. It might enable the sustained processing of information in the stereotypically expected and most frequent way despite encounters of stereotype-violating information. The stereotype might only be updated when these encounters reach a critical amount or are dramatic in nature (Rothbart, 1981). Such a mechanism could help explain why stereotypes are so resistant to change.

# Chapter 3
# Effects of episodic stereotype-relevant representations on the processing of subsequent information

# 6. Goals

The experiments in this chapter investigated the stability and strength of the episodic representations evoked by the reading of stereotype-relevant information by examining its effect on the processing of subsequent information.

In Experiment 1, I found a mismatch effect for the online processing of gender stereotype-relevant information: Stereotype-mismatching reflexive pronouns were looked at longer than matching pronouns. This indicates that participants had detected the inconsistencies. A subsequent cued-recall questionnaire asking for the agents' gender revealed that participants remembered stereotype-relevant information above chance level. This can be taken as an indication that the mismatches had not only been detected but that participants had also -- at least sometimes, temporarily and to some degree -- constructed accurate representations of the stereotype-mismatching information. The conclusions that can be drawn from the results of this questionnaire are, however, limited. There was a time delay between the encoding of the sentences and the memory task and intervening sentences might have interfered with the memory for the representations of particular items. These factors might have caused memory distortions which might account for the significant response bias towards stereotype-matching responses. For the cases in which participants responded incorrectly, it is therefore not clear whether they had initially resolved the mismatches and constructed correct episodic representation, but subsequently forgotten them, or whether they had not constructed stable enough episodic representations in the first place to remember them later.

The goal of the experiments in this chapter was to investigate the nature of the episodic representations evoked by reading stereotype-relevant information by introducing a more immediate way of evaluating them. In Experiment 2 and 3, I examined whether the representations evoked by the reading of stereotype-relevant information were stable and strong enough to generalise from the context of one sentence to the next. As will be explained below, the two experiments differed slightly in the syntactic structure of the sentences: In Experiment 2, the head noun (e.g., *mechanic*) was repeated in the second sentence (as in “In the evening, the *young mechanic* seated *himself*/*herself* comfortably in front of the TV and watched the all-night song contest with an old friend. At bedtime, the *young mechanic* found it difficult to drag *himself*/*herself* away from the program.”), whereas in Experiment 3, reference to the agent was left implicit (as in “In the evening, the young *mechanic* seated *himself/herself* comfortably in front of the TV, watched the all-night song contest with an old friend and, at bedtime, found it difficult to drag *himself/herself* away from the program.”).

# 7. Experiment 2: Effects of episodic stereotype-relevant representations on further processing

## 7.1 Overview

For Experiment 2, I used similar sentences as in Experiment 1 to introduce an agent with a stereotype-relevant occupation coupled with a matching or mismatching pronoun. A second sentence within the same discourse context then repeated the agent and pronoun information. Thus, participants read sentence pairs such as “In the evening, the young *mechanic* seated *himself*/*herself* comfortably in front of the TV and watched the all-night song contest with an old friend. At bedtime, the young *mechanic* found it difficult to drag *himself*/*herself* away from the program”. I used eye-movement measures as reflection of processing difficulty for both the first and second encounter with the agent and pronoun (inclusively spill-over region). Therefore, the results for the first encounter (in sentence 1) provided a point of comparison for the results of the second encounter (in sentence 2). This design allowed for the episodic representations evoked by reading stereotype-relevant information to be examined immediately within the same processing episode.

Apart from examining the influence of episodic representations evoked by a first encounter with an item on the processing of a second encounter with the same item, I also examined its influence on the processing of a subsequent encounter with another item of the same category. I sought to determine whether the representation of a particular stereotype-relevant token (e.g., a specific mechanic who is female) would affect the processing of another instance of this specific token (e.g., the same female mechanic) and generalise to the entire type (i.e., all mechanics). I therefore tested the sentence pairs in two conditions: a *token condition* and a *type condition*. An example for a sentence pair in the token condition is: “In the evening, the *young mechanic* seated *herself/himself* comfortably in front of the TV and watched the all-night song contest with an old friend. At bedtime, the *young mechanic* found it difficult to drag *himself*/*herself* away from the program.” An example for a sentence pair in the type condition is: “In the evening, the *young mechanic* seated *himself*/*herself* comfortably in front of the TV and watched the all-night song contest with an old colleague. At bedtime, the *older mechanic* found it difficult to drag *himself/herself* away from the program.”

## 7.2 Hypotheses

As regions of interest, I defined the agent, pronoun and pronoun spill-over regions in the first and second sentence. As eye-movement measures, I chose First Fixation

Duration, First-Pass Duration, Selective Regression-Path Duration, Total Reading Time, Regression In and Regression Out.

For the eye-movement measures in the first sentence, I expected to replicate the results of Experiment 1: a mismatch effect on the pronoun and pronoun spill-over regions as well as on the late eye-movement measures on the agent region. I did not expect a difference between the token and type conditions, as the first sentences were almost identical in both conditions.

### Measuring the effect of episodic stereotype-relevant representations on the subsequent processing of the same tokens

In the sentence pairs presented in this study, the first sentence constituted a semantic context for the second sentence. There have been previous studies investigating whether increased processing times to gender stereotype-mismatching information can be avoided by introducing disambiguating context information. In their second experiment, Duffy and Keir (2004), had participants read three context sentences before the critical fourth sentence, which included the stereotype-matching or -mismatching occupation-pronoun combination (e.g., “The *electrician* taught *himself/herself* a lot while fixing the problem.”). The second of the context sentences was either gender-disambiguating or neutral. An example for a disambiguating sentence was: “The *electrician* was a cautious *man /woman* who carefully secured *his /her* ladder to the side of the house before checking the roof.” An example for a neutral sentence was: “The *electrician* was cautious and carefully secured the ladder to the side of the house before checking the roof.” Duffy and Keir found a mismatch effect on the pronoun and post-pronoun region in the fourth sentence following neutral but not disambiguating second sentences. This result shows that the mismatch effect can be avoided when a disambiguating context is given.

Carreiras, Garnham, Oakhill and Cain (1996) also investigated the influence of prior disambiguating information on the reading times of stereotype-matching or -mismatching information. They first tested short English text passages that introduced a stereotype-relevant role name in the first sentence and a stereotype-matching or mismatching pronoun in the second sentence. (e.g., "The *footballer* wanted to play in the match. *He/she* had been training very hard during the week."). Carreiras and colleagues found longer total reading times for the second sentences with a mismatching than with a matching pronoun. In a second study, they tested Spanish sentence stimuli that, again, introduced a stereotype-relevant role name in the first sentence (e.g., "*El/la futbolista* quería a jugar el partido", meaning: "The footballer wanted to play in the match"). The difference to the experiment in English, however, was that the role names were gender-disambiguated by the female or male definite articles "la" or "el" and therefore marked from the outset as stereotype-matching or mismatching. After an intervening sentence, the third sentence referred back to the role name with a stereotype-matching or -mismatching pronoun in the third sentence ("*El/Ella* había estado entrenando mucho durante la semana", meaning: "*He/she* had been training hard during the week."). The total reading times showed a stereotype-mismatch effect only in the first but not in the third sentence. The participants appeared to have incorporated the disambiguating information into their episodic representation of the agent. This cancelled a later effect of stereotype-mismatch. Kreiner, Sturt, and Garrod (2008) also showed that a gender-mismatch effect on reading times can be avoided when the gender of a stereotype-relevant agent is disambiguated. They used cataphoric sentences in which the pronoun preceded the role noun (e.g., "After reminding *himself/herself* about the letter, the *minister* immediately went to the meeting at the office."). No mismatch effect was observed on the reading times for the gender-stereotypical agents. If, however, gender-defined role nouns were used (e.g., *the king*), such an effect did occur.

These studies have shown that the occurrence of a gender-mismatch effect on reading times is not inevitable and can be avoided by prior disambiguating context. This focus on avoiding a mismatch effect by presenting disambiguating information differs from the focus of the current experiments. Here, I am interested in the representations constructed during reading sentences including stereotype-matching and -mismatching information, as used in Experiment 1, by examining their effect on further processing. Because of the different foci, the stimuli used in the studies described above differ from those used in the current experiments. Previously, very explicit disambiguating information has been used: Duffy and Keir (2004) used the words *woman* or *man* and in addition the pronouns *her* or *his* to unambiguously describe the agent and Carreiras, Garnham, Oakhill and Cain (1996) used defining definite articles (*la*/*el*) and morphological gender information for the same purpose (e.g., *enfermera*/*enfermero*). In the current study, however, gender was only referred to by stereotype-matching or -mismatching pronouns. As mentioned in section 2, in conceptual terms, pronouns carry meaning only in connection with referents. This makes the gender reference arguably more subtle than the use of nouns referring to gender-specified concepts (Duffy & Keir, 2004) or gender-defined determiners and affixes (Carreiras et al., 1996).

Kreiner, Sturt and Garrod (2008) also used pronouns to supply gender information. In the context of their cataphoric sentences, however, the information the pronouns provided was emphasized by being introduced at the beginning of the sentence, even before the agent. In the present study, I used more subtle disambiguating information than in previous studies and examined the representations evoked by reading gender-stereotype-relevant information and whether they were strong and stable enough to eliminate a second occurrence of the mismatch effect.

Based on my working model, I had certain assumptions about the formation of different representations while participants read the two sentences. When participants encountered the agent region (e.g., *mechanic*) in the first sentence, I assumed that the semantic prototype representation of the agent would be activated. As this representation is linked to the stereotypical gender feature, an expectation would be constructed about the particular agent to have the stereotypical gender (e.g., male). Therefore the stereotypical gender feature would be incorporated in the forming of the episodic representation of the agent. When participants encountered a matching pronoun (e.g., himself) in the first sentence, the episodic representation, including the link to the stereotypical gender feature (e.g., male mechanic), would be confirmed. However, when they encountered a mismatching pronoun (e.g., herself) the episodic representation would be challenged and would need to be changed. It would need to be revised by linking the agent node (e.g., mechanic) to a stereotype-mismatching gender feature node (e.g., female). The extra processing time the mismatch detection and representation update take, should be reflected in the increased viewing times for mismatching compared to matching pronouns. How stable and strong such a newly constructed mismatching exemplar representation is and whether its activation (and therefore accessibility) can outweigh the activation of the prototypical representation in a subsequent processing context has not been investigated before. The episodic representation might be strong and stable enough to be maintained (at least) until the next sentence and to be more accessible than the prototype representation. In this case, the mismatching gender feature linked to the exemplar representation would be confirmed when participants read the mismatching pronoun in the second sentence.

Consequently, a mismatch effect on the second pronoun would be absent (or smaller than on the first pronoun, in case the prototypical gender feature still exerted some influence, too). This would mean that viewing times would be similar for reencountered matching and mismatching pronouns, as well as the respective spill-over and agent regions. Alternatively, the mismatching episodic representations constructed in the first sentence might not be strong and stable enough to override the activation and accessibility of the prototype representation. In this case a mismatch effects would emerge in the second sentence on the pronoun, the pronoun spill-over region, as well as the late measures of the agent region.

### Measuring the effect of episodic stereotype-relevant representations on the subsequent processing of other tokens of the same type

If the episodic exemplar representations constructed after reading the first sentence were maintained and remained active and accessible during the processing of the second sentence, they might not only exert an influence on the further processing of the exact same person (i.e., the specific mechanic), but also generalise to the processing of other persons of the same category (i.e., another mechanic). I tested this by having participants read a second sentence that referred to either the same person as in the first sentence (token condition) or a different person of the same category (type condition). If the processing of a stereotype-mismatching agent could influence the inferences about the gender of another agent to be stereotype-mismatching too, this would have implications for the update and change of stereotypes. In my working model, long-term stereotype change is modelled to happen gradually. Stereotypical prototype representations are abstractions of multiple exemplars and change slowly with new exemplar information. This mechanism resembles the one proposed by Rothbart (1981) in the *bookkeeping model* in which people keep track of stereotype-matching and -mismatching instances and gradually update their stereotypes with each piece of stereotype-mismatching information.  What was tested in this experiment was the *short-term* effect of an activated stereotype-mismatching episodic exemplar representation on the processing of other instances within the same category and, in terms of my working model, on the prototype representation.

It was possible that a mismatching exemplar representation would be treated as a very critical piece of information due to its recency and saliency. It might therefore – in the short-term – outweigh other exemplar information and exert an over-proportional influence on the stereotypical prototype-representation. In this case, when the second agent region was encountered, the salience of the mismatching exemplar representation might exert a strong enough influence on the prototype representation for the activation of the stereotype-mismatching feature to equal or outweigh the activation of the stereotypical gender feature. This could result in the participants' inferences about the second agent's gender to be stereotype-mismatching, too. A mismatching pronoun would be consistent with such an inference, resulting in the absence of a mismatch effect in the second sentence or at least its reduction compared to the first sentence.

It was also possible, however, that a mismatching exemplar representation would – even in the short-term – be treated as just one other exemplar contributing to the abstracted prototype representation without exerting a critical individual influence on the prototype representation. This suggestion is in line with the bookkeeping model. In this case, the encounter with the new agent would activate the stereotypical gender feature linked with the prototype representation. This would result in the incorporation of the stereotype-matching gender-feature in the episodic representation. This representation would be confirmed when a stereotype-matching pronoun would be encountered. It would, however, be challenged when a mismatching pronoun would be encountered. In this case the recognition of the mismatch and the revision of the representation would be expected to result in a repeated mismatch effect for the second sentence.

## 7.3 Method

### Participants

Forty participants completed the experiment (39 female; mean age 19.28 years, ranging 18 to 36 years). All participants were undergraduate or postgraduate students at the University of Birmingham with normal or corrected-to-normal vision and native speakers of British English. They received either course credits or money in exchange for their participation.

### Apparatus

The same apparatus was used as specified in Chapter 2 (see section 5.3).

### Materials

For the reading task, 48 sentence pairs were constructed around the same stereotypically female and male occupations used in Experiment 1. The item *footballer* was replaced by *goalkeeper,* as this experiment required the agents of the sentences to be uniquely identifiable. Whereas there are several football players in a team, there is only one goalkeeper. The item was pretested (amongst five filler item) and analysed in the same way as described in Chapter 2 (see section 5.3; for the pretest questionnaire see Appendix 13). A one-sample *t*-test revealed that the mean scores for female and male typicality were significantly different from the midpoint value of the scales (4). The item *goalkeeper* fulfilled the criterion for a stereotypically male occupation as the mean female typicality rating was greater than 6 (6.60) and the mean male typicality rating was smaller than 2 (1.60).

Half of the 48 sentence pairs were constructed in a way that the agents of the first and second sentence were the same person (token condition; e.g., “In the evening, the *young mechanic* seated *herself/himself* comfortably in front of the TV and watched the all-night song contest with an old friend. At bedtime, the *young mechanic* found it difficult to drag *himself/herself* away from the program.”). The other half of the 48 sentence pairs was constructed in a way that the agents of the first and second sentence were different people with the same occupation label (type condition; e.g., “In the evening, the *young mechanic* seated *himself/herself* comfortably in front of the TV and watched the all-night song contest with an old colleague. At bedtime, the *older mechanic* found it difficult to drag *himself/herself* away from the program.”). For a full listing of all experimental sentences in Experiment 2, see Appendix 14. Both token and type conditions had one version in which the occupation labels were combined with a stereotype-matching reflexive pronoun and another version in which the occupation labels were combined with a stereotype-mismatching reflexive pronoun.

The sentences in the token and type condition were as similar as possible. The sentences of Experiment 1 were adapted as little as possible to fit as first sentences the content in both conditions. The distance between the agent and pronoun regions was kept as similar as possible across conditions.

Each sentence included the same regions as the sentences in Experiment 1: initial region, agent region, reflexive verb region, reflexive pronoun region, pronoun spill-over region, final region. Again, the spill-over region was the word following the pronoun if that word had four or more letters, or two words following the pronoun otherwise. There were a number of other constraints regarding the layout of the sentences. The agent and pronoun regions always appeared on the same line. The agent region was always preceded by at least 10 characters at the beginning of the line. The pronoun region was always followed by at least 10 characters before the end of the line. The spill-over region was always followed by at least 6 characters before the end of the line. The layout of the token and type conditions was matched. Both sentences within one trial were presented in a single paragraph if all the above constraints were fulfilled; otherwise the second sentence was presented in a new paragraph. If this arrangement could still not fulfil all the constraints, the sentences were split up in a way that the constraints were fulfilled. In all cases the sentences in the token and type conditions were broken up in the same way and had the same visual appearance. The sentences were displayed in Courier New, font size 13. One character had an average degree of visual angle of 0.31°.

In order to disguise the purpose of the reading task, 48 filler trials were intermixed randomly with the experimental sentences. Twenty-four of these consisted of three sentences, twelve consisted of two sentences, and twelve consisted of one sentence. Some of the filler sentences served as targets for two other experiments unrelated to the one reported here and had a different linguistic structure than the experimental sentences.

To keep participants alert and assess their overall comprehension, half of all experimental and filler sentences were followed by simple *yes*/*no* comprehension questions. Altogether 38 comprehension questions were presented. *Yes* and *no* push-button responses were recorded.

### Design

The experiment was based on a 2 (token-type: token, type) x 2 (match: match, mismatch) x 2 (sentence number: sentence 1, sentence 2) within-participants design. In the token condition, the first sentence had the same agent as the second sentence. In the type condition, the first sentence had a different agent from the second sentence. As in Experiment 1, 12 of the 24 experimental sentence pairs included a stereotypically female occupation role, and 12 included a stereotypically male occupation role. Six of each of the female and male sentence pairs were combined with stereotype-matching pronouns (match condition), and the other six sentences with stereotype-mismatching pronouns (mismatch condition). The experiment was tested in two versions and orders between participants (see section 5.3 for more details and Appendix 15 for an overview of the design). Comprehension questions required a *yes* or *no* push-button response.

### Procedure

Participants completed a consent form and a short questionnaire specifying age, gender, and first language. Then they read instructions, specifying the requirements of the task (see Appendix 16). The participants' seating position, as well as the calibration, validation and drift correction procedures, were the same as in Experiment 1. The experiment lasted about 35 minutes. After completion, participants were debriefed and received course credits or money.

### Statistical analyses

I used the same analyses methods for the eye-movement measures and comprehension questions as described in Chapter 2 (see section 5.3).

### Analysis of eye-movements

The analyses of the eye-movements followed largely the same rules as specified in Chapter 2, section 5.3. Regions of interest were again the agent region, the pronoun region and the pronoun spill-over region. The entire trial reading time was not included because the sentences in the token and type condition differed slightly to fit the content in both conditions. For example, first sentences in the token condition sometimes introduced a new person differently from first sentences in the type condition where this person acted as the new agent in the second sentence (e.g., "One morning, the beautician spoke aloud to herself/himself about serious family problems without realising that the *receptionist /a young colleague* was listening from the next room."). The second sentences were adapted to make sense in both the token (e.g., "*The unhappy beautician* was deeply ashamed of herself/himself *on learning that the receptionist was gossiping* about these problems.") and type conditions (e.g., "*This recently hired beautician* was deeply ashamed of herself/himself *when caught gossiping* about these problems."). The entire trial reading times in the token and type conditions were therefore not comparable. Also, as first and second sentences were presented within the same trial, separate sentence reading times could not be determined.

Viewing-time measures included First Fixation Duration, First-Pass Duration, Selective Regression-Path Duration and Total Reading Time. Regression-proportion measures included Regression Out and Regression In. The exclusion criteria for the interest region analyses were the same as in Experiment 1 (see 5.3).

## 7.4 Results

First, the results for the comprehension questions will be described. Then, the trial inclusions will be described for each region of interest, followed by the statistical analyses of the dependent measures for each region of interest.

### Comprehension question results

Of the 1520 button-press responses to the comprehension questions 1375 (90%) were correct. A one-sample *t*-test showed that the number of correct responses was significantly above chance ($t(39) = 31.51$, $p < .001$). This result indicates that the participants had read the sentences in the reading task for comprehension, as instructed.

### Trials included in eye data analysis

Table 17 displays the number and percentages of trials included in the eye data analyses for each interest area region after the application of the exclusion criteria.

Table 17: Overview of number (and percentages) of trials included in the eye data analysis for the different interest areas

| | Agent region | Pronoun region | Spill-over region |
|---|---|---|---|
| All trials | 1906 (100%) | 1906 (100%) | 1906 (100%) |
| After blink trials exclusion | 1868 (98.01%) | 1896 (99.48%) | 1884 (98.95%) |
| After skipped trials exclusion | 1682 (88.25%) | 1638 (85.94%) | 1430 (75.03%) |
| After outlier exclusion | 1677 (87.99%) | 1636 (85.83%) | 1424 (74.71%) |

### Statistical analysis

Analyses of variance (ANOVAs) by participants and items were carried out for each interest region and each eye-movement measure with the factors match (match versus mismatch), token-type (token versus type) and sentence (sentence 1 versus sentence 2)[25].

### *The agent region*

For the agent region, I expected a mismatch effect on the later eye-movement measures with longer viewing times and a larger proportion of regressions for mismatching than matching sentences. Depending on whether the exemplar representation constructed in the first sentence was strong and stable enough to influence the processing of the second sentence, these effects were expected to be smaller or absent in the second sentence. This was expected to possibly result in an interaction between match and sentence in the token condition. In case such an interaction would be found in the token condition, a generalisation to the type condition might also occur.

The agent region means for the different eye-movement measures can be found in Tables 18 and 19. A table with the cell means and standard deviations used in the ANOVA participant analyses can be found in Appendix 17. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 18.

A mismatch effect was observed only for the late eye-movement measure Regression In, significant when analysed by participants ($F_1(1,39) = 4.14$, $p < .05$) and marginally significant when analysed by items ($F_2(1,23) = 3.26$, $p = .084$). This means that after participants had moved on in the sentence, they were more likely to look back into the agent region when the sentence had turned out to be mismatching than when it was matching.

[25] The factor half was not included as it did not yield many theoretically interesting effects in Experiment 1.

There were sentence effects on all viewing-time measures with longer viewing times in the first than in the second sentence: First Fixation Duration ($F_1(1,39) = 7.66$, $p < .01$; $F_2(1,23) = 5.65$, $p < .05$), First-Pass Duration ($F_1(1,39) = 38.80$, $p < .001$; $F_2(1,23) = 41.45$, $p < .001$), Selective Regressive Path Duration ($F_1(1,39) = 54.77$, $p < .001$; $F_2(1,23) = 44.64$, $p < .001$), Total Reading Time ($F_1(1,39) = 58.26$, $p < .001$; $F_2(1,23) = 52.40$, $p < .001$). The Regression In measure shows a larger proportion of regressions in the first than in the second sentence ($F_1(1,39) = 5.18$, $p < .05$; $F_2(1,23) = 6.74$, $p < .05$). These effects indicate that participants spent more time gazing at the agent region when they encountered it in the first sentence than in the second sentence. There were, however, no significant interactions between match and sentence.

No difference was observed between the token and type conditions. There was, however, a Token-Type x Sentence interaction for the Selective Regression-Path measure, significant by participants and nearly significant by items ($F_1(1,39) = 4.70$, $p < .05$; $F_2(1,23) = 4.19$, $p = .052$). Post-test analyses revealed that the difference in sentence viewing times showed the same direction in both conditions. In the token condition the viewing time was longer for the first sentence than in the second sentence (361 msec versus 317 msec; $F_1(1,39) = 19.10$, $p < .001$; $F_2(1,23) = 38.65$, $p < .001$). The same was true for the type condition (380 msec versus 303 msec; $F_1(1,39) = 38.16$, $p < .001$; $F_2(1,23) = 27.64$, $p < .001$). The difference between the viewing times on the first and second sentences, however, was greater in the type (77 msec) than in the token (43 msec) condition.

No interaction was observed between match and sentence in the token or type conditions, indicating that the processing of the first sentence did not affect the processing of the agent in the second sentence, regardless of whether it was the same agent or a different one than in the first sentence.

Table 18: Eye-movement measure means for the agent region by sentence

| | Sentence 1 | | | Sentence 2 | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-match | Total | Match | Mis-match | Total |
| First Fix. Duration | 247 | 248 | 247 | 238 | 236 | 237 | 243 | 242 | 242 |
| First-Pass Duration | 342 | 355 | 349 | 298 | 293 | 296 | 320 | 324 | 322 |
| Selective Reg.-Path Duration | 364 | 377 | 371 | 315 | 306 | 310 | 340 | 341 | 340 |
| Total Reading Time | 442 | 474 | 458 | 362 | 362 | 362 | 402 | 418 | 410 |
| Regression Out | 0.10 | 0.11 | 0.11 | 0.12 | 0.12 | 0.12 | 0.11 | 0.12 | 0.11 |
| Regression In | 0.12 | 0.18 | 0.15 | 0.10 | 0.11 | 0.10 | 0.11 | 0.14 | 0.13 |

Table 19: Eye-movement measure means for the agent region by token-type

| | Token | | | Type | | |
|---|---|---|---|---|---|---|
| | Match | Mismatch | Total | Match | Mismatch | Total |
| First Fix. Duration | 242 | 242 | 242 | 243 | 242 | 242 |
| First-Pass Duration | 321 | 321 | 321 | 320 | 327 | 324 |
| Selective Reg.-Path Duration | 341 | 337 | 339 | 338 | 345 | 342 |
| Total Reading Time | 399 | 415 | 407 | 405 | 422 | 413 |
| Regression Out | 0.11 | 0.11 | 0.11 | 0.11 | 0.12 | 0.11 |
| Regression In | 0.11 | 0.17 | 0.14 | 0.11 | 0.12 | 0.12 |

***The pronoun region***

For the pronoun region, I expected a mismatch effect on all eye-movement measures with longer viewing times and a larger proportion of regressions for mismatching than matching sentences. If a sufficiently strong and stable episodic representation was constructed in the first sentence to influence the processing of the second sentence, an interaction between match and sentence was expected with smaller or absent mismatch effects in the second sentence. In case of such an interaction in the token condition, a generalisation to the type condition might also occur.

The pronoun region means for the different eye-movement measures can be found in Tables 20 and 21. A table with the cell means and standard deviations used in the ANOVA participant analyses can be found in Appendix 19. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 20.

As expected, mismatch effects were observed for all eye-movement measures with longer fixation times and a larger numbers of fixations in the mismatch than in the match condition: First Fixation Duration ($F_1(1,39) = 13.31$, $p < .005$; $F_2(1,23) = 5.51$, $p < .05$), First-Pass Duration ($F_1(1,39) = 28.36$, $p < .001$; $F_2(1,23) = 7.72$, $p < .05$) , Selective Regressive Path Duration ($F_1(1,39) = 42.93$, $p < .001$; $F_2(1,23) = 13.92$, $p < .005$) , Total Reading Time ($F_1(1,39) = 54.43$, $p < .001$; $F_2(1,23) = 21.61$, $p < .001$), Regression Out ($F_1(1,39) = 6.11$, $p < .05$; $F_2(1,23) = 5.42$, $p < .05$), Regression In ($F_1(1,39) = 4.21$, $p < .05$; $F_2(1,23) = 4.71$, $p < .05$).

There was a consistent pattern for all viewing-time measures of longer gazes on the pronoun in the first sentence than in the second sentence. This difference reached significance in the Total Reading Time, analysed by participants ($F_1(1,39) = 5.79$, $p < .05$) and marginal significance, analysed by items ($F_2(1,23) = 3.32$, $p = .082$).

There was, however no interaction between match and sentence.

No main effects were observed for the token-type factor. There was, however, an interaction between token-type and match in the Regression Out measure ($F_1$(1,39) = 4.21, $p$ < .05; $F_2$(1,23) = 5.22, $p$ < .05). The same was true for the Regression In measure for which, however, the interaction was only marginally significant in the participant analyses ($F_1$(1,39) = 3.47, $p$ = .07; $F_2$(1,23) = 6.18, $p$ < .05). Post-hoc test showed that the interactions were due to the fact that the proportion of regressions was significantly greater in the mismatch condition than in the match condition only in the token (Regression Out: $F_1$(1,39) = 9.20; , $p$ < .001; $F_2$(1,23) = 13.19, $p$ < .005; Regression In: $F_1$(1,39) = 10.22, $p$ < .005: $F_2$(1,23) = 9.66, $p$ < .01), but not in the type condition (Regression Out: $F_1$(1,39) = .03; , $p$ = .868; $F_2$(1,23) = .01, $p$ = .924; Regression In: $F_1$(1,39) = .01, $p$ = .923: $F_2$(1,23) = .02, $p$ = .891). No a priori expectations had been formulated about such interactions. The interpretation of these results is limited, as the regression-proportion measures contain information about other regions of the sentences from or to which regressions were made. These regions might have differed slightly between the token and type conditions. I had, however, formulated hypotheses about the interaction of match and sentence in the token and type conditions. The occurrence of such an interaction in the token condition would have indicated that the processing of the first sentence affected the processing of the pronoun in the second sentence. In case such an interaction occurred, it was expected to potentially generalise to the type condition. As, however, no such interaction was observed in the token condition, it was unsurprisingly also not observed in the type condition.

Table 20: Eye-movement measure means for the pronoun region by sentence

| | Sentence 1 | | | Sentence 2 | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-Match | Total | Match | Mis-match | Total |
| First Fix. Duration | 216 | 228 | 222 | 213 | 224 | 219 | 215 | 226 | 220 |
| First-Pass Duration | 235 | 257 | 246 | 225 | 243 | 234 | 230 | 250 | 240 |
| Selective Reg.-Path Duration | 241 | 274 | 258 | 237 | 259 | 248 | 239 | 267 | 253 |
| Total Reading Time | 304 | 355 | 329 | 280 | 328 | 304 | 292 | 341 | 316 |
| Regression Out | 0.07 | 0.14 | 0.10 | 0.10 | 0.11 | 0.10 | 0.09 | 0.12 | 0.10 |
| Regression In | 0.13 | 0.16 | 0.14 | 0.12 | 0.17 | 0.14 | 0.13 | 0.16 | 0.14 |

Table 21: Eye-movement measure means for the pronoun region by token-type

| | Token | | | Type | | |
|---|---|---|---|---|---|---|
| | Match | Mismatch | Total | Match | Mismatch | Total |
| First Fix. Duration | 215 | 227 | 221 | 214 | 226 | 220 |
| First-Pass Duration | 233 | 251 | 242 | 227 | 249 | 238 |
| Selective Reg.-Path Duration | 241 | 269 | 255 | 237 | 265 | 251 |
| Total Reading Time | 286 | 338 | 312 | 298 | 344 | 321 |
| Regression Out | 0.07 | 0.14 | 0.11 | 0.10 | 0.10 | 0.10 |
| Regression In | 0.11 | 0.18 | 0.14 | 0.15 | 0.15 | 0.15 |

### *The pronoun spill-over region*

The hypotheses for the pronoun spill-over region were derived from the expectations for the pronoun region: an effect of match on all eye-movement measures with longer viewing times and a larger proportion of regressions for mismatching than matching sentences and potentially an interaction between match and sentence with smaller or absent mismatch effects in the second sentence. In case of such an interaction, a generalisation from the token condition to the type condition might also occur.

The pronoun spill-over region means for the different eye-movement measures can be found in Tables 22 and 23. A table with the cell means and standard deviations used in the ANOVA can be found in Appendix 21. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 22.

Consistent with my expectations, mismatch effects were observed fo all viewing-time measures with longer fixation times in the mismatch than in the match condition: First Fixation Duration ($F_1(1,39) = 7.34$, $p < .05$; $F_2(1,23) = 5.58$, $p < .05$), First-Pass Duration ($F_1(1,39) = 4.56$, $p < .05$; $F_2(1,23) = 9.48$, $p < .01$) , Selective Regression-Path Duration ($F_1(1,39) = 11.01$, $p < .005$; $F_2(1,23) = 8.72$, $p < .01$) , Total Reading Time ($F_1(1,39) = 15.58$, $p < .001$; $F_2(1,23) = 19.48$, $p < .001$).

Significantly longer viewing times were observed for the first sentence than the second sentence in the early viewing-time measures: First Fixation Duration ($F_1(1,39) = 5.31$, $p < .05$; $F_2(1,23) = 10.48$, $p = .005$) and First-Pass Duration ($F_1(1,39) = 4.25$, $p < .05$; $F_2(1,23) = 5.50$, $p = .05$). As for the pronoun region, no interaction between match and sentence was observed.

No main effects were observed for the token-type factor. There were, however, interactions between token-type and sentence in the late viewing-time measures Selective Regression-Path measure ($F_1(1,39) = 7.36$, $p < .05$; $F_2(1,23) = 6.57$, $p < .05$) and Total Reading Time ($F_1(1,39) = 11.95$, $p < .005$; $F_2(1,23) = 6.00$, $p < .05$). No a priori expectations had been formulated about such an interaction and no post-hoc test were carried out as the sentence stimuli were slightly different in the token and type conditions, making an interpretation difficult.

Table 22: Eye-movement measure means for the pronoun spill-over region by sentence

| | Sentence 1 | | | Sentence 2 | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-Match | Total | Match | Mis-match | Total |
| First Fix. Duration | 228 | 240 | 234 | 219 | 227 | 223 | 223 | 233 | 228 |
| First-Pass Duration | 274 | 296 | 285 | 264 | 275 | 269 | 269 | 285 | 277 |
| Selective Reg.-Path Duration | 294 | 325 | 309 | 288 | 306 | 297 | 291 | 316 | 303 |
| Total Reading Time | 350 | 405 | 377 | 344 | 381 | 363 | 347 | 393 | 370 |
| Regression Out | 0.12 | 0.16 | 0.14 | 0.17 | 0.17 | 0.17 | 0.15 | 0.17 | 0.16 |
| Regression In | 0.12 | 0.17 | 0.14 | 0.16 | 0.17 | 0.16 | 0.14 | 0.17 | 0.15 |

Table 23: Eye-movement measure means for the pronoun spill-over region by token-type

| | Token | | | Type | | |
|---|---|---|---|---|---|---|
| | Match | Mismatch | Total | Match | Mismatch | Total |
| First Fix. Duration | 227 | 240 | 234 | 219 | 226 | 223 |
| First-Pass Duration | 268 | 295 | 281 | 270 | 276 | 273 |
| Selective Reg.-Path Duration | 285 | 321 | 303 | 296 | 310 | 303 |
| Total Reading Time | 351 | 388 | 370 | 343 | 398 | 371 |
| Regression Out | 0.13 | 0.16 | 0.14 | 0.16 | 0.18 | 0.17 |
| Regression In | 0.18 | 0.16 | 0.17 | 0.09 | 0.18 | 0.13 |

**Summary of eye-movement results**

A summary of the match and sentence results for viewing time and proportion regression measures for the agent, pronoun, and pronoun spill-over regions can be found in Table 24. None of the results of the type and token conditions differed significantly from each other and therefore the token-type factor is omitted from this table.

Table 24: Summary of eye-movement results for the factors match and sentence (sen)

| | Agent region | | | Pronoun region | | | Pronoun spill-over region | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Sen | Match x Sen | Match | Sen | Match x Sen | Match | Sen | Match x Sen |
| First Fix. Duration | - | X | - | X | - | - | X | X | - |
| First-Pass Duration | - | X | - | X | - | - | X | X | - |
| Selective Reg.-Path Duration | - | X | - | X | - | - | X | - | - |
| Total Reading Time | - | X | - | X | - | - | X | - | - |
| Regression Out | - | - | - | X | - | - | - | - | - |
| Regression In | - | X | - | X | - | - | - | - | - |

## 7.5 Discussion

The goal of Experiment 2 was to assess the stability and strength of episodic representations evoked by reading stereotype-relevant information. This was done by examining their influence on the subsequent processing of sentences referring either to the same person (token condition) or a different person of the same category (type condition). The results of Experiment 2 showed a mismatch effect with longer viewing times and a larger proportion of regression for the stereotype-mismatching pronouns and pronoun spill-over regions. However, no interactions were observed between the factors match and sentence for neither the token nor the type condition. The repeated mismatch effect in the second sentence might have emerged because the mismatching episodic representations formed in the first sentence were not strong and stable enough be maintained until the second sentence and to outweigh the activation and accessibility of the prototype representation. This would mean that on encountering the second pronoun, the matching gender feature linked to the prototype representation was more accessible than the mismatching gender feature. The possibility that a mismatching episodic representation would not even have an effect on the further processing within the same processing episode, however, seems to be at odds with the results of the memory questionnaire in Experiment 1. They had shown that participants were – at least sometimes – able to remember stereotype-mismatching information after a time delay of up to 30 minutes and the presentation of several intervening items. An alternative explanation is that after participants had constructed mismatching episodic representations in the first sentence, the activation of the prototype representation was boosted again when at the beginning of the second sentences the occupation label was mentioned again. Consequently, the prototypical, matching gender-feature would have been more activated than the episodic, mismatching one. On encountering the second mismatching pronoun, this would have lead to a repeated mismatch detection and effort for an update of the episodic representation.

In order to investigate these possibilities and to clarify the source of the repeated mismatch effect in Experiment 2, I designed Experiment 3. Here, I adjusted the syntactic structure of the sentences of the token condition to exclude a second mention of the occupation label. The type condition was not considered in Experiment

3, as the reformulation of the sentences without inclusion of a second occupation label was only possible for the token condition.

# 8. Experiment 3: Effects of removing repeated category labels on subsequent processing

## 8.1 Overview

In Experiment 3, I sought to disambiguate the source of the mismatch effect in the second sentence of Experiment 2. If it had occurred because the occupation label at the beginning of the second sentence had reactivated the gender stereotype, it should disappear when such a label was missing. I therefore rephrased the sentences of the token condition of Experiment 2 so that the occupation label was mentioned only once at the beginning. I did this by connecting the two sentences to one. This only worked for the sentence pairs in which the text that intervened between the two pronouns did not introduce a new sentential subject. This was the case for 16 of the 24 sentence pairs in Experiment 2. I analysed the results of Experiment 2 separately for this subset and the results and effects pattern remained the same as for the entire sample.

In this experiment, I not only investigated the effect of the representations evoked by reading stereotype-relevant information on the processing of information within the same processing episode, but also the effects on longer-term memory performance. To this end, I appended a memory questionnaire measure after the reading task[26]. This questionnaire addressed a few methodological limitations of the questionnaire in Experiment 1. Specifically, the questionnaire in Experiment 1 did not distinguish between items participants had and had not seen in the reading task, but rather between stereotype-matching and -mismatching responses (e.g., "Was the secretary

[26] With hindsight it would have been useful to administer this questionnaire in Experiment 2 as well.

male/female?”). This meant that these two types of responses were not independent of each other and could not be analysed separately. The questionnaire in Experiment 3, however, was designed as a recognition test for items participants had encountered in the reading task (*old items*) and items participants had not encountered in the reading task (*new items*). This meant that matching and mismatching responses were independent of each other and could be assessed separately. This was particularly useful, as I was specifically interested in the long-term representation of the mismatching items. Another improvement was that the way information was to be recalled was more similar to the way it had been encoded in the new compared to the old questionnaire. In Experiment 1, the bare role names had been used as questionnaire items, whereas in Experiment 3, sentences were used. The old items were the sentences exactly as presented in the reading task (e.g., stereotype-matching) and the new items were the sentences as presented in the reading task, but with a different pronoun (e.g., stereotype-mismatching). I presented the sentences one by one on the screen just as in the reading task, unlike the paper and pencil version, used in Experiment 1.

## 8.2 Hypotheses

### Measuring the effects of removing repeated category labels on subsequent processing

If the repeated mismatch effect on the second pronoun emerged because the episodic representations constructed for the first sentence were insufficiently strong and stable to successfully compete against the stereotypical prototype representation, the results of Experiment 3 should replicate the results of Experiment 2. In this case, I expected mismatch effects for both the first and second pronoun without an interaction between match and pronoun number. Alternatively, if the repeated mismatch effect emerged because the repetition of the agent's occupation label gave the prototype representation an additional activation boost over the episodic representation, the mismatch effect should be confined to the first pronoun. I expected effects on the pronoun regions to extend to the pronoun-spill-over regions - particularly the early measures. There was only one agent region per trial, for which I did not expect mismatch effects in any of the early measures as the pronoun had not yet been encountered. There had been no effects of match on the later eye-movement measures on the agent region in Experiment 2. Based on this result, I did not expect an effect on these measures in Experiment 3. Late effects on this region would, however, be in line with the idea that participants look back into the region more after encountering an mismatching than a matching pronoun to ensure that they had read correctly and possibly also to aid the process of inconsistency resolution.

### Measuring the effects of episodic, stereotype-relevant representations on memory

Regarding the memory questionnaire, I had separate hypotheses for memory sensitivity and bias. The questionnaire in Experiment 1 showed that participants were able to remember stereotype-relevant information above chance level. The stimuli in the present questionnaire had been improved in terms of similarity between the items in the reading task and the items in the questionnaire. I therefore expected memory sensitivity to again be significantly above chance for both matching and mismatching items. Stangor and McMillan's (1992) meta-analyses of memory for expectancy-congruent and -incongruent information showed that recognition sensitivity was generally greater for expectancy-mismatching information. Therefore memory sensitivity in the present study might be better for mismatching than matching items, too.

Participants in Experiment 1 had shown a conservative bias toward stereotype-matching responses. Therefore, for this questionnaire too, I expected that participants would fall back onto their gender stereotypes as cues when they could not remember an agent’s gender. There was, however, the possibility that participants would rely less on their stereotypes when trying to remember the items in this experiment than in Experiment 1. Here, the pronoun was repeated in the reading task, possibly leading to stronger and more stable representations and less conservative memory bias. In Stangor and McMillan’s (1992) meta-analyses, response bias was generally stronger for expectancy-matching information. It was therefore possible that a similar result might be found in this experiment.

## 8.3 Methods

### Participants

Thirty-two participants completed the experiment (24 female; mean age 20.94 years, ranging 18 to 28 years). All participants were undergraduate or postgraduate students at the University of Birmingham with normal or corrected-to-normal vision and native speakers of British English. They received either course credits or money in exchange for their participation.

### Apparatus

The same apparatus was used as specified in Chapter 2 (see section 5.3).

## Materials

### *Reading task materials*

For the reading task the sentence stimuli were adapted from the token condition in Experiment 2. The sentences were reformulated so that they did not repeat the occupation label. I did this by connecting each former sentence pair to a single sentence. This was possible if the intervening text did not introduce a new sentential subject. A subset of sixteen items fulfilled this condition. Overall, the sentences were changed as little as possible in comparison to Experiment 2. For a full listing of all experimental sentences in Experiment 3 see Appendix 23.

The constraints for the layout of the sentences were as described in Experiment 2. The same 48 filler items were used as in Experiment 2. The eight experimental sentence pairs from Experiment 2 that could not be combined into one sentence also served as fillers. The same 36 comprehension questions were used as in Experiment 2 (see section 7.3).

### *Memory task materials*

The memory task consisted of the 24 stereotype-relevant sentences from the reading task. Of these, the sixteen sentences that had been reformulated into one-sentence paragraphs were the target items, the remaining items were fillers.

## Design

In order to make this experiment comparable to Experiment 2, all 24 former experimental items were presented in the same design (i.e., order and match conditions) as in Experiment 2. As only 16 of these items were in the new target subset, the number of stereotypically female and male occupation labels was not counterbalanced. Instead there were seven stereotypically female and nine stereotypically male agents in the first version of the target set and nine stereotypically female and seven stereotypically male agents in the second version. This was not considered a problem as an inspection of the data in Experiment 1 showed similar effects for both stereotypically female and male occupation labels. Therefore, the factor gender had not been included in any of the analyses. Furthermore, the number of matching and mismatching trials was not counterbalanced within one experimental version. Across the two versions, however, match was counterbalanced. In the first version, five of the stereotypically female occupation labels were paired with a matching pronoun and four with a mismatching pronoun. Five of the stereotypically male occupation labels were paired with a matching pronoun and two with a mismatching pronoun. In the second version all sentences that were matching in the first version were mismatching, and all sentences that were mismatching in the first version were matching. As in Experiment 2, both experimental versions were presented in two different orders.

The memory questionnaire had two versions, corresponding to the reading task version. Half the items in each version matched the items in the reading task version and half mismatched them. According to the two orders of the reading task versions, the memory questionnaire versions also had two different orders.

**Procedure**

Participants completed a consent form and a short questionnaire specifying age, gender, and first language. Then they read instructions, specifying the requirements of the task (see Appendix 16). The participants’ seating position, as well as the calibration, validation and drift correction procedures, were the same as in

Experiment 1. Participants carried out the reading task first and subsequently the surprise memory task. The memory task required participants to indicate whether they recognised a sentence as the exact same sentence as they had read in the reading task. They were informed that if a sentence would be different, it would only differ in the agent's gender (see Appendix 24). The memory task was self-paced and participants indicated their response by pressing the *yes* button on the response box if they thought the sentence was exactly the same as in the reading task, the *no* button if they thought the sentence differed in the agent's gender.

The experiment lasted approximately 30 minutes. After completion, participants were debriefed and received course credits or money.

### Statistical analyses

I used the same analyses methods for the eye-movement measures and comprehension questions as described in Chapter 2 (section 5.3). For the analysis of the memory questionnaire, I applied signal detection theory and performed log-linear analyses.

### Analyses of eye-movements

The analyses of the eye-movements followed the same procedure, rules and criteria as specified in section 7.3.

## 8.4 Results

### Reading task results

#### *Comprehension Questions*

Of the 1216 responses to the comprehension questions 1098 (90%) were correct. A one-sample *t*-test showed that the number of correct responses was significantly above chance ($t(31) = 28.98$, $p < .001$). This result indicates that the participants had read the sentences in the reading task for comprehension, as instructed.

***Trials included in analyses of eye movements***

Table 25 displays the number and percentages of trials included in the eye data analyses for each interest area region after the application of the exclusion criteria.

Table 25: Overview of number (and percentages) of trials included in eye data analysis for the different interest areas

| | Agent region | Pronoun region | Spill-over region |
|---|---|---|---|
| All trials | 512 (100%) | 1024 (100%) | 1024 (100%) |
| After blink trials exclusion | 507 (99.01%) | 1020 (99.61%) | 1012 (98.83%) |
| After skipped trials exclusion | 485 (94.73%) | 873 (85.25%) | 821 (80.18%) |
| After outlier exclusion | 485 (94.73%) | 870 (84.96%) | 820 (80.08%) |

***Statistical analysis***

Analyses of variance (ANOVAs) with participants and items as random variables were carried out for each interest region and each eye-movement measure. For the agent, there was only one factor (match), for the pronoun and spill-over regions there were two crossed factors of match (match versus mismatch) and pronoun number (pronoun1 versus pronoun 2).

*The agent region*

For the agent region, I had expected neither early effects nor, based on the results of Experiment 2, late effects. The agent region means for the different eye-movement measures can be found in Table 26. A table with the cell means and standard deviations used in the ANOVA participant analyses can be found in Appendix 25. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 26.

There was a pattern in the later viewing-time measures and the late Regression In measure of longer viewing times and greater proportions of regressions in the mismatch than in the match condition. Although none of these differences were significant, the pattern supports the assumption that participants looked back into the agent region more often after encountering a mismatching than a matching pronoun. This probably serves the purpose of checking whether they had read correctly and might help resolve the stereotype-mismatch.

Table 26: Eye-movement measure means for the agent region by match

| | Match | Mismatch |
|---|---|---|
| First Fixation Duration | 225 | 221 |
| First-Pass Duration | 318 | 335 |
| Selective Reg.-Path Duration | 346 | 357 |
| Total Reading Time | 460 | 498 |
| Regression Out | 0.13 | 0.11 |
| Regression In | 0.18 | 0.23 |

*The pronoun region*

I had expected that if the episodic representation participants had constructed in the first sentence was strong and stable enough to outweigh the influence of the prototype representation on the processing of the second sentence, a mismatch effect for the pronoun region and an interaction with pronoun number would emerge. The mismatch effect for the second pronoun was in this case expected to be smaller than the mismatch effect for the first pronoun or absent. If the episodic representation was less activated and accessible than the prototype representation in the second sentence, in contrast, I had expected a repeated mismatch effect on the second pronoun. The pronoun region means for the different eye-movement measures can be found in Table 27. A table with the cell means and standard deviations used in the ANOVA participant analyses can be found in Appendix 27. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 28.

Mismatch effects were observed for the later viewing-time measures Selective Regressive Path Duration ($F_1(1,31) = 6.82$, $p < .05$; $F_2(1,15) = 6.37$, $p < .05$) and Total Reading Time ($F_1(1,31) = 7.74$, $p < .01$; $F_2(1,15) = 7.33$, $p < .05$), as well as for the late regression-proportion measure Regression In ($F_1(1,31) = 9.24$, $p < .01$; $F_2(1,15) = 4.97$, $p < .05$). The earlier regression-proportion measure Regression Out was significant by items only ($F_1(1,31) = 2.54$, $p = .121$; $F_2(1,15) = 5.16$, $p < .05$). For all these measures, longer fixation times and a larger number of fixations were observed in the mismatch than in the match condition.

There was a consistent pattern for all eye-movement measures of significantly longer gazes and larger proportions of regressions for the first than the second pronoun: First Fixation Duration ($F_1(1,31) = 17.96$, $p < .001$; $F_2(1,15) = 9.01$, $p < .01$), First-Pass Duration ($F_1(1,31) = 14.53$, $p < .005$; $F_2(1,15) = 7.01$, $p < .05$) , Selective Regressive Path Duration ($F_1(1,31) = 20.85$, $p < .001$; $F_2(1,15) = 12.73$, $p < .005$) , Total Reading Time ($F_1(1,31) = 29.52$, $p < .001$; $F_2(1,15) = 12.81$, $p < .005$), Regression Out ($F_1(1,31) = 4.25$, $p < .05$; $F_2(1,15) = 5.91$, $p < .05$), Regression In ($F_1(1,31) = 9.17$, $p < .01$; $F_2(1,15) = 4.30$, $p = .056$). These results suggest that it took the participants less effort to process the second pronoun than the first pronoun. This is not surprising, given that the second pronoun repeated the information that had the participants had already processed when they read the first pronoun.

There were significant interactions between the factors match and pronoun number on the late viewing-time measures Selective Regression-Path Duration ($F_1(1,31) = 7.43$,

$p < .05$; $F_2(1,15) = 7.23$, $p < .05$) and Total Reading Time ($F_1(1,31) = 25.26$, $p < .001$; $F_2(1,15) = 34.19$, $p < .001$) as well as the late regression-proportion measure Regression In ($F_1(1,31) = 6.17$, $p < .05$; $F_2(1,15) = 7.25$, $p < .05$). The interaction was also marginally significant on the earlier regression-proportion measure Regression Out ($F_1(1,31) = 3.57$, $p = .068$; $F_2(1,15) = 4.34$, $p = .055$).

Planned comparisons revealed that for the following measures the interactions were due to significantly longer viewing times and larger number of regressions in the mismatch than in the match condition on the first pronoun only: Selective Regression-Path Duration ($F_1(1,31) = 9.52$, $p < .01$; $F_2(1,15) = 11.90$, $p = .005$), Regression Out ($F_1(1,31) = 4.49$, $p < .05$; $F_2(1,15) = 6.03$, $p < .05$), Regression In ($F_1(1,31) = 13.04$, $p = .005$; $F_2(1,15) = 8.58$, $p = .05$). The interaction on the Total Reading Time was due to significantly longer viewing times in the mismatch than in the match condition on the first pronoun ($F_1(1,31) = 21.03$, $p < .001$; $F_2(1,15) = 22.05$, $p = .001$), but marginally longer viewing times in the match than in the mismatch condition on the second pronoun ($F_1(1,31) = 3.34$, $p = .077$; $F_2(1,15) = 4.51$, $p = .051$).

The significant interactions of match and pronoun number on the late viewing-time measures and regression-proportion measures, revealing the absence of a mismatch effect on the second pronoun, suggest that participants had constructed strong and stable enough representations after an encounter with the first mismatching pronoun to influence the further processing of stereotype-relevant information. This account is supported by the results in the early viewing-time measures: Although the interactions did not reach significance, the result pattern is the same as for the late measures.

Table 27: Eye-movement measure means for the pronoun region by pronoun number

| | Pronoun1 | | | Pronoun 2 | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-match | Total | Match | Mis-match | Total |
| First Fix. Duration | 216 | 223 | 219 | 203 | 201 | 202 | 209 | 212 | 211 |
| First-Pass Duration | 231 | 247 | 239 | 215 | 216 | 215 | 223 | 231 | 227 |
| Selective Reg.-Path Duration | 236 | 277 | 256 | 220 | 220 | 220 | 228 | 248 | 238 |
| Total Reading Time | 303 | 409 | 356 | 295 | 267 | 281 | 299 | 338 | 319 |
| Regression Out | 0.07 | 0.15 | 0.11 | 0.06 | 0.06 | 0.06 | 0.06 | 0.11 | 0.08 |
| Regression In | 0.14 | 0.28 | 0.21 | 0.13 | 0.14 | 0.14 | 0.13 | 0.21 | 0.17 |

*The pronoun spill-over region*

I had expected the effects of the pronoun region to possibly extend to the pronoun spill-over region: a mismatch effect with or without an interaction between the factors match and pronoun number, depending on whether the episodic representation constructed after reading the first pronoun outweighed in activation and accessibility the prototype representation.

Pronoun spill-over region means for the different eye-movement measures can be found in Table 28. A table with the cell means and standard deviations used in the ANOVA can be found in Appendix 29. A table with the complete ANOVA results for each eye-movement measure can be found in Appendix 30.

There was a consistent pattern of longer gazes in the mismatch than the match condition for all viewing-time measures. None of these differences, however, reached significance. The Regression Out measure showed a greater proportion of regressions in the mismatch than in the match condition. This difference was significant when analysed by participants and almost significant when analysed by items: $F_1(1,31) = 5.48$, $p < .05$; $F_2(1,15) = 4.40$, $p = .053$.

No main effects were observed for the factor pronoun number. For the Selective Regression-Path Duration, there was an interaction, marginally significant analysed by items, but not by participants ($F_1(1,31) = 2.07$, $p = .160$; $F_2(1,15) = 3.94$, $p = .066$). As planned comparisons showed, it was due to significantly longer viewing times in the mismatch than match condition on the first pronoun only ($F_1(1,31) = 5.98$, $p < .05$; $F_2(1,15) = 7.12$, $p < .05$). There was also an interaction on the Regression Out measure, significant by participants, but not by items ($F_1(1,31) = 6.93$, $p < .05$; $F_2(1,15) = 3.09$, $p = .099$), which again was due a mismatch effect on the first but not the second pronoun. For the first pronoun, the proportion of regressions made out of the region into earlier parts of the sentence was significantly greater in the mismatch than in the match condition ($F_1(1,31) = 11.33$, $p < .005$; $F_2(1,15) = 7.80$, $p < .05$).

In sum, the post-hoc analyses for the pronoun spill-over regions revealed a mismatch effect in the first but not the second spill-over region for the Selective Regression-Path Duration and the Regression Out measures. These measures are closely associated with the processing of the pronoun region: At least parts of the regressions made out of the spill-over region can be assumed to be made into the pronoun region; the Selective Regression-Path Duration reflects the time spent on the spill-over region before it is processed enough to carry on reading. This processing time includes the durations of fixations made after shifts to the left of the region for rereading. Again, at least some of the content that was reread can be assumed to be the pronoun. Hence, the fact that these measures showed a larger proportion of regressions and longer viewing time, respectively, for the mismatching than the matching sentences only in the first, but not the second pronoun, supports the hypotheses that participant had, after reading the first pronoun, constructed strong and stable enough representations to affect the processing of the second pronoun.

Table 28: Eye-movement measure means for the pronoun spill-over region by pronoun number

| | Pronoun 1 | | | Pronoun 2 | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Match | Mis-match | Total | Match | Mis-Match | Total | Match | Mis-match | Total |
| First Fix. Duration | 228 | 232 | 230 | 220 | 218 | 219 | 224 | 225 | 225 |
| First-Pass Duration | 280 | 296 | 288 | 277 | 277 | 277 | 278 | 287 | 282 |
| Selective Reg.-Path Duration | 300 | 342 | 321 | 310 | 307 | 309 | 305 | 325 | 315 |
| Total Reading Time | 399 | 427 | 413 | 388 | 400 | 394 | 393 | 414 | 403 |
| Regression Out | 0.11 | 0.23 | 0.17 | 0.14 | 0.14 | 0.14 | 0.13 | 0.19 | 0.16 |
| Regression In | 0.18 | 0.14 | 0.16 | 0.16 | 0.16 | 0.16 | 0.17 | 0.15 | 0.16 |

***Summary of eye-movement results***

A summary of the results for viewing time and proportion regression measures for the agent, pronoun, and pronoun spill-over regions can be found in Table 29.

Table 29: Summary of significant eye-movement results for the factors match and pronoun number

| | Agent region | Pronoun region | | | Pronoun spill-over region | | |
|---|---|---|---|---|---|---|---|
| | Match | Match | Pronoun number | Match x Pronoun number | Match | Pronoun number | Match x Pronoun number |
| First Fix. Duration | - | - | X | - | - | - | - |
| First-Pass Duration | - | - | X | - | - | - | - |
| Selective Reg.-Path Duration | - | X | X | X | - | - | - |
| Total Reading Time | - | X | X | X | - | - | - |
| Regression Out | - | - | X | - | X | - | - |
| Regression In | - | X | X[27] | X | - | - | - |

**Memory task results**

I had expected participants' responses in the questionnaire to depend on the information they had encoded during the reading task, and therefore memory sensitivity to be above chance for both matching and mismatching sentences. Based on previous findings (Stangor & Mcmillan, 1992), I had further considered it possible that the sensitivity measure might favour mismatching information. Regarding memory bias, I had expected that participants would show a conservative response

[27] Marginal by items (p = .056)

bias towards stereotype-matching sentences and that the bias measure might be stronger for matching than for mismatching sentences.

The number of *old* and *new* responses to old or new stimuli in the questionnaire for both matching and mismatching questionnaire sentences can be found in Table 30. Correct responses are marked with (c), incorrect responses with (i).

A signal detection theory analyses revealed a memory sensitivity of $d'_{match}$= 1.23 for the matching sentences and $d'_{mismatch}$ = 0.96 for the mismatching sentences. The response bias was found to be $C_{match}$ = -0.46 for the matching, and $C_{mismatch}$ = -0.08 for the mismatching sentences, indicating a conservative response bias in both conditions (for the procedure of calculating memory sensitivity and bias see Chapter 2, section 5.3).

In order to determine whether sensitivity and bias were significantly different from 0 and whether the differences between matching and mismatching sentences were significant, I carried out a log-linear analysis. A hierarchical fully saturated log-linear analysis was applied to the present data set with the factors item (old, new), response (old, new)[28] and match (match, mismatch). The interactions included the two-way interactions Item x Response[29], Item x Match and Response x Match and the three-way interaction Item x Response x Match.

The log-linear analyses revealed that the appropriate model of the data did not require the three-way Item x Response x Match interaction ($G^2(1) = 1.66$; $p = .20$) or the two-way Item x Match interaction ($G^2(1) = 1.44$; $p = .23$). It did, however, require the two-way Item x Response interaction ($G^2(1) = 86.39$; $p < .001$) and the two-way Response x Match interaction ($G^2(1) = 9.21$; $p < .005$). These interactions were included in the

[28] This is another formulation for response bias

[29] This is another formulation for memory sensitivity

final formulation of the model along with the factors that yielded a main effect: item, response and match.

That the Item x Response interaction was a required factor of the model shows that participants' responses depended on which stimuli they had seen (compare 357 correct responses with 155 incorrect responses in Table 30). As the interaction with the factor match was not required for a model of the data, memory sensitivity did not differ significantly between the matching and mismatching sentences.

The result of response being a required factor of the model shows that there was a significant response bias. The descriptive data showed that the bias was conservative, leaning towards expectancy consistency in both match conditions. The result that the Response x Match interaction was a required factor for a model shows that the bias was affected by consistency. As has been seen in the descriptive data, bias was greater for the matching than the mismatching sentences ($C_{match}$ = -0.46 versus $C_{mismatch}$ = -0.08).

Overall, the results of the questionnaire show that, as expected, participants were able to distinguish sentences they had read from sentences they had not read well above chance level, independently of whether the sentences had been matching or mismatching. This indicates that the representations of the agents were long-lasting enough to be accessed or reconstructed later on. If participants were unsure or not able to remember, however, they were more inclined to indicate that they had seen a sentence before when the questionnaire sentence was matching rather than mismatching. This was consistent with my expectations and indicates that they activated and consulted their gender-stereotype when recollection was missing.

Table 30: Number and percentage of correct and incorrect *old* and *new* responses to old and new questionnaire items

| | Matching sentences | | Mismatching sentences | | Total | |
|---|---|---|---|---|---|---|
| | Response: 'old' | Response: 'new' | Response: 'old' | Response: 'new' | Response: 'old' | Response: 'new' |
| Old items | 110 (c) | 18 (i) | 91 (c) | 37 (i) | 201 (c) | 55 (i) |
| New items | 56 (i) | 72 (c) | 44 (i) | 84 (c) | 100 (i) | 156 (c) |
| SUM | 166 (65%) | 90 (35%) | 135 (53%) | 121 (47%) | 301 (59%) | 211 (41%) |

## 8.5 Discussion

The goals of Experiment 3 had been to disambiguate the source of the repeated mismatch effect on the second pronoun in Experiment 2. I also sought to investigate the effect of these representations on longer-term memory.

I found an interaction between match and pronoun number with a mismatch effect for the first, but not the second pronoun. By contrast, in Experiment 2, I had found a mismatch effect on both pronouns. This pattern indicates that the repeated mismatch effect in Experiment 2 was due to a reactivation of the stereotype by a repetition of the category label.

The results of the recognition questionnaire showed that participants were able to remember both stereotype-matching and -mismatching information above chance. This suggests that the representations constructed when the agent and pronoun information was first introduced did not only have a short-term effect on the further processing within the same processing episode, but could also be accessed again later. When participants could not remember whether they had seen a sentence, they seem to have used their gender stereotype as guessing aid, as suggested by the conservative response bias, In line with the results by Stangor and McMillan (1992), this bias was stronger for matching than mismatching items.

# 9. General discussion

## 9.1 Discussion of the online findings

### Summary of the findings

In Experiment 2, participants read sentence pairs about agents with stereotype-relevant occupations. An agent was introduced with an occupation label at the beginning of the first sentence (e.g., *mechanic*) and referred to again by a matching (e.g., *himself*) or mismatching (e.g., *herself*) reflexive pronoun. At the beginning of the second sentence, the occupation label was repeated, referring to either the same agent (token condition) or to a different agent (type condition), followed again by a matching or mismatching reflexive pronoun. The eye-movement measures in this experiment revealed mismatch effects in both sentences, with longer viewing times and larger number of regressions for the stereotype-mismatching than the stereotype-matching pronouns and pronoun spill-over regions. The token-type manipulation did not reveal any processing differences and was not followed up in the next experiment. If an interaction had been found between match and sentence in the token condition, I had expected that a generalisation might occur to the type condition. As, however, such an interaction was not found in the token condition, no conclusions can be drawn about such a generalisation. Therefore, it has to remain unresolved whether a single encounter with a stereotype-mismatching agent can – at least short-term – affect the inferences about the gender of another agent to be stereotype-mismatching, too. Experiment 3 sought to clarify whether the repeated mismatch effect in the second sentence of Experiment 2 arose because the episodic representations constructed in the first sentence were not strong and stable enough to outweigh the influence of the stereotypical prototype representation on processing the subsequent information, or whether the repeated mismatch effect arose due to a reactivation of the stereotypical prototype representation after the second encounter with the occupation label. To disambiguate the source of the second mismatch effect, the sentences were reformulated in order to avoid a second mention of the occupation label. With the adapted sentences, a mismatch effect was only observed on the first, but not the second pronoun.

### Interpretation of the findings within the working model

These results can be interpreted in terms of the working model. When participants encountered the first agent, the prototype representation with a link to the stereotype-matching gender node was activated. Therefore the initial episodic representation also had a link to a stereotype-matching gender node. If the first pronoun was matching, this representation was confirmed. If the pronoun was mismatching, this mismatch was detected and the episodic exemplar representation had to be updated with a new link to a stereotype-mismatching node. The mismatch detection and representation update was reflected in longer viewing times for mismatching than matching pronouns. As no mismatch effect was found for the second pronoun in Experiment 3, the exemplar representation is assumed to have been strong and stable enough to be maintained and remain active until the encounter with that region. In Experiment 2, however, the agents' category membership was reemphasised by a repetition of the occupation label. The repeated occupation label seems to have given rise to a reactivation of the semantic prototype representation including the stereotypical gender link, as, when the second pronoun was encountered, a repeated mismatch effect on the second pronoun emerged.

**Placing the findings within the context of prior research**

Previous studies have demonstrated that mismatch effects did not emerge when prior disambiguating information was presented. Duffy and Keir (2004) showed this with paragraphs such as: “Jeff ’s/Lucy’s power had been unreliable ever since the tornado. The *electrician* was a cautious *woman/man* who carefully secured *her/his* ladder to the side of the house before checking the roof. Jeff/Lucy suspected that high winds had loosened the connection to the power lines. The *electrician* taught *herself/himself* a lot while fixing the problem.” (p. 555). As can be seen, Duffy and Keir introduced an agent with a stereotypical occupation (e.g., *electrician*) as woman or man in the second of four sentences. Within the same sentence they referred to the agent again with a pronoun (*her/his*). They found an initial mismatch effect on the viewing times on the *woman* or *man* regions. However, in the fourth (target) sentence they did not find such an effect on the pronoun. Like in Experiment 2, the occupation was repeated at the beginning of the target sentence but, unlike in Experiment 2, this did not cause a second mismatch effect on the pronoun. This could be because the explicit and repeated gender information might have given rise to a stronger episodic exemplar representation than in Experiment 2. As a consequence, the exemplar representation might have not been outweighed by the semantic prototype representation, as assumed for Experiment 2.

Carreiras, Garnham, Oakhill and Cain (1996) tested English sentences such as “The *footballer* wanted to play in the match. *He/she* had been training very hard during the week” and Spanish sentences like “*El/la futbolista* quería a jugar el partido.

[intervening second sentence] *El/Ella* había estado entrenando mucho durante la semana" (these example sentences have the same meaning as the English sentences). In the English sentences, the gender of the agent was left ambiguous until the pronoun at the beginning of the second sentence. Carreiras and colleagues found a mismatch effect for the total reading time for this sentence, suggesting that participants had constructed a stereotype-matching episodic representation. In the Spanish sentences, the gender of the agent was disambiguated by a definite article (*el/la*) and for some of the stimulus sentences also by the morphological form of the suffix (e.g., *enfermero/enfermera*). Carreireas et al. found a mismatch effect for the first, but not the third (target) sentence. This suggests that participants had constructed an episodic representation based on the morphosyntactic rather than the stereotypical gender information.

Both Duffy and Keir and Carreiras et al. used more explicit gender disambiguating information than I did in my experiments. Perhaps the use of such explicit information changes the saliency of the gender information. This difference in saliency could explain why the second mention of the occupation role in the experiment by Duffy and Keir did not reactivate the stereotype enough to cause a second mismatch effect. It could be that the episodic representation that was initially formed of a female electrician included a very strongly salient gender node (*female*). When the occupation label was mentioned a second time, it activated the semantic occupation node with a link to the stereotypical gender node (*male*). However, because of the saliency of the female node linked to the episodic representation, its activation would not be outweighed by the activation of the stereotypical gender node. In my Experiment 2, the gender information about the agent was subtly only referred to with a pronoun: "After work, the plumber got *herself* a big portion of chips even though the doctor had strongly recommended a low-fat diet. The hungry plumber was unable to control *herself* when it came to chips.” This might have lead to the formation of an episodic representation of the female plumber with a less salient stereotype-mismatching gender node than in the experiment by Duffy and Keir. When the second occupation region was entered, the stereotype-mismatching gender node of the episodic representation might therefore have been outweighed by the stereotype-matching gender node of the semantic prototype representation. The suggestion that the saliency of the gender information might change depending on how information is presented, is consistent with the results by Kreiner, Sturt and Garrod (2008). They presented occupation- and gender-relevant information in two different orders. In anaphora sentences, the occupation information preceded the gender information:”Yesterday the *minister* left London after reminding *himself/herself* about the letter”. In cataphoric sentences, the gender information preceded the occupation information: “After reminding *himself/herself* about the letter, the *minister* immediately went to the meeting at the office”. A mismatch effect was only observed in the anaphoric sentences. Here, the gender feature would have initially been informed by the stereotypical prototype representation and then updated when a mismatching pronoun was encountered. The mismatch detection and resolution are the processes assumed to be reflected in the mismatch effect. In the cataphora sentences, the episodic representation was first informed by the gender node. When readers encountered the occupation node, it got incorporated in the representation without causing a mismatch effect.

In sum, the suggestion that the way and order in which written stereotype-relevant information is presented might lead to differences in the salience of gender information within the episodic representational could explain different findings regarding the effect of prior disambiguating information on stereotype-mismatch effects.

## 9.2 Discussion of the memory findings

### Summary of the findings

In Experiment 3, participants carried out a memory task examining the longer-term stability of the episodic representations constructed during the reading task. The task repeated the sentences of the reading task with either the same pronoun as before (old items, requiring an *old* response) or a different pronoun than before (new items, requiring a *new* response). For the analyses, I compared the number of *old* and *new* responses to old and new items separately for stereotype-matching and -mismatching questionnaire sentences. Table 31 clarifies the relationship of the stereotype-matching and –mismatching sentences in the reading task to stereotype-matching and -mismatching sentences in the memory task in terms of the corresponding correct responses.

Table 31: The relationship of sentences in the reading and memory task regarding correct responses

| Sentence in the reading task | Stereotype-matching | | Stereotype-mismatching | |
|---|---|---|---|---|
| Sentence in the memory task | Stereotype-matching | Stereotype-mismatching | Stereotype-mismatching | Stereotype-matching |
| Correct response | Old | New | Old | New |

### Placing the findings within the context of prior research

The sensitivity findings showed that participants recognised both originally matching and mismatching sentences correctly at above chance levels. This indicates that the episodic representations formed when reading the sentences were stable enough not only to avoid the reappearance of the mismatch effect in the same processing episode, but also to be retrieved or reconstructed later. Memory bias was conservative towards ‘old’ responses for both matching and mismatching questionnaire sentences, but stronger for matching ones. The bias findings indicate that when participants could not retrieve or reconstruct the episodic representations, they used the semantic representations as guessing aid. These results are consistent with the source of activation confusion (SAC) model of memory (see Figure 3; Diana, Reder, Arndt, & Park, 2006; Reder, Nhouyvanisvong, Schunn, Ayers, Angstadt, & Hiraki, 2000). Within the SAC model a *concept node* (corresponding to the semantic prototype representation in the current working model) is activated when a word is encoded during an experiment. The general experimental environment (e.g., room temperature, lighting) is represented by an *experimental context node*. The specific encoding environment of the current trial (e.g., participants’ reaction to a stimulus, incidental noise) is represented by a *specific context node*. The concept, the experimental and specific context nodes are all linked to an *episode node* (corresponding to the episodic exemplar representation in the current working model) that is constructed when a word is encoded.

When item recognition is required after the experiment, recollection and familiarity processes have been distinguished. Recollection was suggested to correspond to participants’ ability to *remember* specific details about encoding a particular item during the experiment, in other words, the activation of the episode node. Familiarity was suggested to correspond to participants not recalling specific details about encoding a particular item during the experiment, but, on the basis of familiarity with the item, still somehow having the feeling of *knowing* it, in other words, the activation of the concept node. To tap into these processes, participants have in many memory experiments been explicitly asked to not only indicate whether or not they had seen an item, but also whether they made their decision on basis of remembering or knowing (for a meta-analysis see Gardiner, Ramponi, & Richardson-Klavehn, 2002).

The bias finding in the present study might rely on the same processes outlined in the SAC model. When participants, during the memory task, tried to recognise whether they had encoded a particular stereotype-matching sentence during the reading task, they might have remembered specific details about studying the item and that it had been stereotype-matching, corresponding to the activation of the episodic exemplar node. In this case, they would have given a stereotype-matching response, based on recollection (resulting in a correct *old* response to stereotype-matching questionnaire sentences or a correct *new* response to stereotype-mismatching questionnaire sentences). If, however, they could not remember specific details of encoding the item and/or whether it was stereotype-matching, they might have, nevertheless, had a feeling of knowing that, for example, the mechanic, was male. This would correspond to the activation of the semantic prototype node and have lead participants to give a stereotype-matching response based on familiarity (again, resulting in a correct *old* response to stereotype-matching questionnaire sentences or a correct *new* response to stereotype-mismatching questionnaire sentences). When participants tried to recognise whether they had encoded a particular stereotype-mismatching sentence during the reading task, they, again, might have remembered specific details about studying the item and that it had been stereotype-mismatching. This would correspond to the activation of the episodic exemplar node and participants would have given a stereotype-mismatching response, based on recollection (resulting in a correct *old* response to stereotype-mismatching questionnaire sentences or a correct *new* response to stereotype-matching questionnaire sentences). If, however, they could not remember specific details of encoding the item and/or whether it was stereotype-mismatching, the stereotype-matching prototype representation might be activated, leading to a stereotype-matching response (resulting in an incorrect *old* response to stereotype-matching questionnaire sentences or an incorrect *new* response to stereotype-mismatching questionnaire sentences). Within this framework, response bias would be conservative for both stereotype-matching questionnaire items (regardless of being encoded or not) and -mismatching questionnaire items (matching response is here false alarm). It naturally would be stronger for matching questionnaire sentences, however, as processes both of recollection (associated with the activation of the episodic representation) and familiarity (associated with the activation of the prototype representation) would favour 'old' responses. For mismatching questionnaire sentences, only the process of recollection would favour 'old' responses.

The interpretation based on the SAC model is also congruent with the Encoding Flexibility Model by Sherman, Lee, Bessenoff and Frost (1998) that predicts that stereotypes not only allow perceivers to allocate attention efficiently, but also to reconstruct memory when it fails.

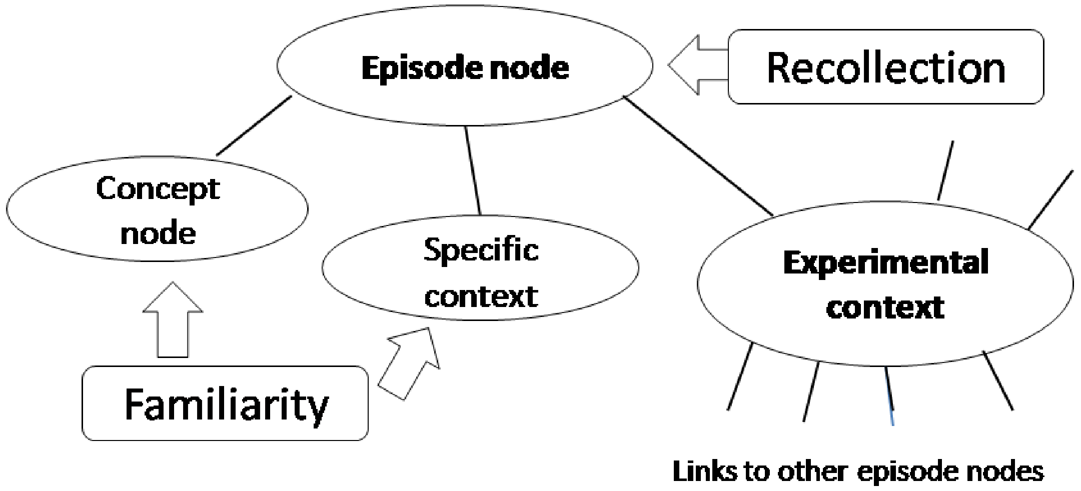


Figure 3: Schematic representation information storage in memory according to the source of activation confusion (SAC) model of memory (adapted from Diana, Reder, Arndt & Park, 2006).

## 9.3 Implications of the findings for stereotype change

In sum, the findings of Experiments 2 and 3 indicate that written stereotype-mismatching information about a particular member of a stereotype-relevant group can be retained within the same processing episode unless the stereotype is reactivated by repeatedly highlighting the stereotype-group label. Also, stereotype-mismatching information can be recognised later on above chance level, even though memory is biased towards stereotype-matching information. These results suggest that readers can construct stereotype-mismatching exemplar representations that are stable and long-lasting enough to influence the further immediate processing and to be recollected later. In my working model, semantic prototype representations are abstractions of the sum of relevant exemplar representations, which means that when enough stereotype-mismatching exemplars are encoded, there is a chance of the stereotype to be adjusted and updated. Unfortunately, the implications of the findings regarding particular members of a stereotype-relevant group for other members remain inconclusive, as the token-type manipulation in Experiment 2 did not render any informative results.

# Chapter 4
# Effects of stereotype-relevant linguistic context on face processing

# 10. Goals

In Chapter 3, I describe evidence that written stereotype-relevant information had an influence on further linguistic processing. Within the working model described in section 3, the explanation for this finding was that episodic representations were constructed when readers first encountered occupation labels and stereotype-matching or -mismatching pronouns, which governed the further processing of the text. In the present chapter, I investigated the scope of these episodic representations as interpretative frameworks for the processing of new information. I examined whether the influence on further processing would be limited to linguistic information or could also extend cross-modally to non-linguistic information, specifically pictorial information. I therefore combined the reading task with a probe task featuring a picture of female and male faces.

Elsewhere, studies that have used mixed sentence and picture stimuli have sought to establish whether semantic processing was modal or amodal (Clark, 1987; Federmeier & Kutas, 2001; Kroll, 1990; Potter & Kroll, 1987; Potter, Kroll, Yachzel, Carpenter, & Sherman, 1986). Potter and colleagues (1986), for example, used a task in which participants processed either regular sentences or sentences in which the critical word was replaced by a picture. Their main result that “rebus sentences were only marginally more difficult to understand and remember than equivalent all-word sentences” (p. 291) led the authors to suggest modality-unspecific conceptual processing. Potter and Kroll (1987) argue that such a conceptual coding model “has greater explanatory power and is more parsimonious than the dual-coding model put forward by Clark [1987] as an alternative” (p. 311) to account for the findings with modality-specific verbal and imaginal components of meaning.

Kroll (1990) used a task in which sentences ended with either a word or non-word about which participants made lexical decisions, or a picture of an object (or non-object) about which participants made object decisions. Reaction times did not differ significantly for word and picture targets. Kroll argues that these results suggest “that words and pictures access common conceptual representations” (p. 753).

Federmeier and Kutas (2001) tested sentences ending with either an expected or unexpected word or picture and found similar ERP responses to words and pictures. They therefore argue that “in line with the predictions of common code models, the organisation of the semantic knowledge store that is accessed by pictures and words seems to be basically similar” (p. 221).

In sum, previous findings suggest that words and pictures share amodal conceptual representations (but see Paivio, 1971, 1986). Within my working model, I share this assumption. Based on this assumption, I had certain expectations about the influence of the episodic representation constructed during reading sentences containing stereotype-relevant information on female or male faces. This influence was assessed by comparing participants’ response times when gender-categorising the face (e.g., as female) after reading a sentence with an agent that matched the face in terms of gender (e.g., “Last week the *secretary* familiarised herself with the new photocopier”) or that mismatched the face in terms of gender (e.g., “In the evening the *mechanic* seated himself comfortably in front of the TV”). If the episodic representation affects the processing of the pictorial information, participants should be faster to categorise a face that matched rather than mismatched the stereotypical gender of the agent. Note that the effect of gender match or mismatch between an agent and a pronoun was not investigated in the experiments included in this chapter. Instead, only gender-matching pronouns were presented.

In the experimental sentences used so far, the gender-marked pronouns always occurred towards the end of the sentences following the agents. Apart from its influence on the episodic representation of the agent, the pronoun might have a conceptual priming effect on the picture in its own right. English pronouns are gender-specific and therefore activate a male or female gender feature at the conceptual level. Because the reflexive pronouns were always the most recent gender-relevant information participants read about before encountering the pictures, they could have a crucial effect on the picture processing. In order to examine whether this was the case, I compared the reaction times for picture-categorisation following sentences with matching agent-pronoun combinations (e.g., “In the evening the *mechanic* seated *himself* comfortably in front of the TV”) to the reaction times following sentences that were similar in meaning but did not include a reflexive pronoun (“In the evening the *mechanic* sat down comfortably in front of the TV”). Unfortunately, the results of the first experiment in this chapter, Experiment 4, were inconclusive. I therefore ran two further experiments with improved designs. The results of these experiments did not reveal any differences between trials with sentences with or without pronouns. More importantly, they did not suggest any influence of episodic stereotype-relevant representations constructed during sentence reading on the processing of pictorial information. In the final experiment, Experiment 7, I therefore investigated whether a cross-modal effect of reading stereotype-relevant information on processing pictorial information would emerge with simpler stimuli, namely, bare gender-stereotypical nouns.

# 11. Experiment 4: Effects of stereotype processing during sentence reading on face processing

## 11.1 Overview

In Experiment 4, I tested the influence of the episodic representations constructed by reading gender-stereotype-relevant information on the further processing of female and male faces. I also tested whether there would be a difference between the influence of sentences that included a stereotype-relevant occupation label and a stereotype-matching reflexive pronoun (*pronoun condition*; e.g., “Last week the *secretary* familiarised *herself* with the new photocopier”) and sentences that included only a stereotype-relevant occupation label (*no-pronoun condition*; e.g., “Last week the *secretary* became familiarised with the new photocopier”).

## 11.2 Hypotheses

In this experiment, a gender-categorisation task with pictures of female and male faces was preceded by linguistic stereotype-relevant information. Other experiments have used similar methodologies to examine questions relevant to social psychology. Kawakami and Dovidio (2001) used gender-stereotypic trait words as primes (e.g., “caring”, “technical”) and photographs of female and male faces as targets. They found that gender-categorisation for the photographs was facilitated when the pictures and preceding words were gender-matched compared to gender-mismatched. Lemm, Dabady and Banaji (2005) used occupation labels that were stereotypically associated with a gender (e.g., “mechanic”, “hairdresser”) or morphologically gender-marked (e.g., “congressman”, “congresswoman”) as primes and line drawings of women and men as targets to assess whether social category knowledge was automatically activated when such words are encountered. Control primes were labels of professions that were equally likely to be associated with men or women (e.g., “author”, “student”) and gender-neutrally suffixed words (e.g., “chairperson”, “salesperson”).

For female targets, Lemm and colleagues found facilitation of picture gender-categorisation times when the primes were gender-matching (compared to male and neutral targets). For male targets, they found an interference effect on picture gender-categorisation times when the primes were gender-mismatching (compared to female and neutral targets). These findings show that cross-modal facilitation effects can be obtained from linguistic stereotype-relevant information to non-linguistic stimuli. These studies used single word primes. The priming effects demonstrate the activation of stereotypical conceptual knowledge by words strongly associated with the stereotypes (Lemm et al., 2005). By contrast, Experiment 4 tested whether more elaborate episodic representations during sentence reading provided an interpretative framework for new information.

Based on my working model, I expected that when participants encountered the occupation label (e.g., *secretary*) during reading, the prototype representation and its associated gender node (*female*) should be activated and consequently also be part of the newly constructed episodic representation. When participants encountered the picture, a similar process should take place. If a picture of a woman is encountered, the prototypical conceptual representation of a woman should be activated. Part of that representation should be a strong link to the *female* gender node. If the picture of a man is encountered, the prototypical conceptual representation of a man with its associated *male* gender node should be activated. These gender features should consequently be part of the episodic representation of the pictures. When the occupation label and the picture match in gender—for example, when both are female—the activation of the *female* node associated with the occupation label should facilitate the activation of the *female* node associated with the *woman* representation evoked by the picture. I expected this to facilitate the categorisation latencies for matching pictures in comparison to mismatching pictures.

In the sentences including a gender-stereotype-matching pronoun, I expected an additional facilitation effect on the picture categorisation latencies in the pronoun compared to the no-pronoun condition. This might be because the pronoun contributes to the generation of the episodic representation by confirming and reactivating the stereotype-matching gender feature. This might make the gender feature more salient within the episodic representation. The additional facilitation effect might also be due to the pronoun activating the same gender feature as the noun at the conceptual level. As most recent gender information before the faces, they might therefore facilitate picture categorisation times.

I included picture gender as a factor in the analysis because previously differential priming effects have been found for female and male pictures (Lemm et al., 2005). However, Lemm and colleagues point out that “This asymmetrical effect may have occurred because the feminine primes were more strongly associated with femininity than the masculine primes were with masculinity. This interpretation is consistent with the finding that explicit ratings of the prime words were not symmetrical” (p. 227). In the present study, however, the pretest ratings for the stereotypical female and male occupation labels were symmetrical (see section 5.3). I therefore did not expect differential effects of sentences including stereotypically female and male occupation labels on picture categorisation latencies.

In sum, I expected the picture-categorisation latencies to be shorter when a face matched rather than mismatched the occupation label in gender. I further expected an interaction between the match and pronoun conditions with a stronger priming effect in the pronoun than in the no-pronoun condition.

## 11.3 Method

### Participants

Twenty-eight participants completed the experiment (all female; mean age 19.32 years, ranging 18 to 25 years). They were randomly assigned to the four versions of the materials in equal numbers. All participants were undergraduate or postgraduate students at the University of Birmingham with normal or corrected-to-normal vision and native speakers of British English. They received course credits or money in exchange for their participation.

### Apparatus

The stimuli were presented on a 17" Samsung Syncmaster 793s monitor controlled by a Javelin computer (Windows XP) running MediaLab and DirectRT research software (Empirisoft Corporation, 2006). Experimenter-specified keys of a standard English keyboard were used as response buttons.

### Materials

#### *Reading materials*

I tested the picture categorisation task in two sentence conditions: the pronoun condition (e.g., “Last week the secretary familiarised herself with the new photocopier”) and the no-pronoun condition (e.g., “Last week the secretary became familiarised with the new photocopier”). The sentences in the pronoun condition were the same sentences as used in the matching condition of Experiment 1; the pronoun always matched the gender stereotype. The sentences in the no-pronoun condition were similar in meaning but did not include the reflexive pronouns (for a full listing of all stimuli see Appendix 31). Each condition comprised 24 sentences. The same 12 female and 12 male sentential subjects were used in both conditions.

To conceal the purpose of the study, forty-eight filler sentences were used (e.g., "The green car pulled out too early and a yellow car had to stop"). The filler sentences were mostly about objects. When people were mentioned, the referents were gender-neutral (e.g., pupils, children). Four additional practice sentences fulfilled the same criteria as the filler sentences. All sentences were presented centrally in black on a white screen with the font face New Courier, size 20.

To ensure that participants would attend to the sentences and to assess overall comprehension, half of the filler sentences were followed by simple *yes*/*no* comprehension questions (e.g., "Did the yellow car stop?"). Altogether, 24 comprehension questions were presented. *Yes* and *no* push-button responses were recorded.

### ***Picture materials***

For the experimental trials, 24 colour pictures were used, 12 depicting female and 12 depicting male faces. For the filler trials, another set of 24 pictures was used, again 12 depicting female and 12 depicting male faces. For the four practice trials, two female and two male faces were used. All pictures were of young Caucasians with a neutral facial expression. The pictures were 640 pixels wide and 480 pixels high.

## **Design**

The experiment was based on a 2 (picture: female, male), 2 (sentence: pronoun, no-pronoun) x 2 (match: match, mismatch) within-participants design. The corresponding sentences in the pronoun and no-pronoun conditions were combined with the same photos. That means that the sentence about the secretary, for instance, was always combined with the same picture, regardless of whether it appeared with or without a pronoun. In the match condition, the gender of the face matched that of the sentential subject; in the mismatch condition, the gender of the face mismatched that of the sentential subject. Sentence-photo pairings were counterbalanced across the four conditions. Each agent appeared once in the pronoun and once in the no-pronoun condition, and once in the matching condition and once in the mismatching condition. The stimuli were assembled into four different versions with 24 experimental trials each with 6 trials from each condition. Half of the six sentences had female and half had male sentential subjects. During the course of the experiment, each participant read one sentence about each agent, for instance a secretary, and saw each face once. The 48 filler trials were mixed in with the experimental trials. Half of the fillers were followed by a photo, and half of the fillers were followed by a comprehension question. The same filler sentences in combination with the same pictures and comprehension questions were tested in the same fixed randomised order in all four versions of the experiment. The experimental items were put in the same slots between the fillers in all four versions with no two experimental trials following each other in direct succession. The first block of the experiment started with four practice trials.

**Procedure**

Participants read the instructions on the screen (see Appendix 32). They were asked to move through the sentences on their own pace by pressing the key on their keyboard with the *Y* or the key with the *N* sticker and to respond to the comprehension questions by pressing the key with the *Y* sticker (*yes* response) or the key with the *N* sticker (*no* response). To categorise the pictures, they were asked to press the key with the *F* sticker for *female* and the key with the *M* sticker for *male*. The *Y* sticker was attached to the *X* key, the *N* sticker was attached to the *Z* key, the F sticker was attached to the "." key, and the *M* sticker was attached to the "/" key on the keyboard. This arrangement meant that the participants used index and middle fingers of their left hand to respond to the comprehension questions and to move on from the sentences, and the index and middle fingers of their right hand to categorise the pictures. All parts of the experiment were self-paced.

The experiment consisted of three blocks with 24 experimental and filler sentences in each. The first block additionally included four practice sentences at the beginning: two followed by a picture, two followed by a question. Between blocks, participants could take short breaks. After the experiment, participants completed a computerised questionnaire specifying their age, handedness, sex, and first language. The experiment lasted approximately 15 minutes. After completion, the experimenter thanked and debriefed the participants.

## 11.4 Results

### Comprehension question results

Participants responded correctly to 613 of 672 (91%) comprehension questions. A one-sample *t*-test showed that the number of correct responses was significantly above chance ($t(27) = 27.11$, $p < .001$). This result indicates that participants read the sentences for comprehension.

**Picture categorisation results**

Participants responded correctly to 97% of the 1344 trials of the picture categorisation task. Responses were correct for 652 of 672 experimental trials and 649 of 672 filler trials. One-sample *t*-tests showed that the number of correct responses was significantly above chance for both the experimental trials ($t(27) = 63.72$, $p < .001$) and filler trials ($t(27) = 88.33$, $p < .001$), indicating that participants completed the picture categorisation task as instructed. Incorrect trials were excluded from the analyses in all experiments in this chapter.

I analysed the categorisation latencies for experimental trials only. I had hypothesised that participants would be faster to categorise faces presented after sentences whose sentential subject matched versus mismatched the face in gender. I further expected an interaction between the match and pronoun conditions with a stronger priming effect in the pronoun than in the no-pronoun condition. I did not expect any differences between processing times for female versus male faces.

The mean categorisation latencies for male and female pictures following a gender-matching versus -mismatching sentence in the pronoun and no-pronoun conditions are presented in Table 32. The data did not contain any extreme outlier values to be excluded from the analyses[30]. It is common in the picture processing literature to only report analyses by participants (e.g., Le Gal & Bruce, 2002; Quinn & Macrae, 2005; Quinn, Mason, & Macrae, 2009; Rossion, 2002). The results of these analyses are reported here ($F_1$). However, following the suggestion by Clark (1973) to not only use

[30] In this and the other experiments in this chapter, fixed upper limits were used to determine outliers (see Ratcliff, 1993) to be consistent with the analyses described in Chapters 2 and 3. However, analyses were also carried out excluding reaction times which were outside a range of the subject mean +/- 2.5 standard deviations (e.g., Quinn & Macrae, 2005). These analyses led to the same conclusions as those reported here.

participants but also items as random variable, I also reported the results of the item analyses ($F_2$) for all experiments. A table with the complete ANOVA results for Experiments 4 to 7 can be found in Appendix 36.

A 2 (picture: male, female) x 2 (sentence: pronoun, no-pronoun) x 2 (match: match, mismatch) within-participants ANOVA revealed that there was no difference between the latencies in the match versus mismatch conditions ($F_1$(1, 27) = .01, *p* = .91) or for the female versus male pictures ($F_1$(1, 27) = 3.29, *p* = .081). It showed further that there was a significant difference between the pronoun and no-pronoun conditions ($F_1$(1, 27) = 5.79, *p* < .05), with faster reaction times in the no-pronoun than in the pronoun condition (761 msec versus 789 msec). The interaction between the sentence and picture factors and the interaction between the match and sentence factors were not significant: ($F_1$(1, 27) = 0.54, *p* = .47; $F_1$(1, 27) = 3.21, *p* = .084). The interaction between the match and picture factors and the interaction between the match, sentence and picture factors were, however, significant ($F_1$(1, 27) = 5.24, *p* < .05; $F_1$(1, 27) = 7.70, *p* < .05).

As Table 32 shows, the reaction times were shorter in the match than in the mismatch condition in the female no-pronoun and pronoun conditions and in the male no-pronoun condition. In contrast, the reaction times were slower in the match than in the mismatch condition in the male pronoun condition.

Separate analyses were conducted for the female and male pictures to examine in which conditions the match and mismatch conditions differed. For the female pictures, no significant differences were found for the sentence factor ($F_1$(1, 27) = 1.44, *p* = .24), the match factor ($F_1$(1, 27) = 2.97, *p* = .096) or the interaction ($F_1$(1, 27) = 0.42, *p* = .53). For the male pictures, no significant differences were found for the sentence factor ($F_1$(1, 27) = 3.33, $p$ = .079) and the match factor ($F_1$(1, 27) = 2.90, $p$ = .10). The interaction, however, was significant: $F_1$(1, 27) = 10.03, $p$ < .005. Analyses of simple effects showed that there was no difference between the match and mismatch condition ($F_1$(1, 27) =1.16, $p$ = .29) in the no-pronoun condition, but there was a significant difference between the match and mismatch condition in the pronoun condition ($F_1$(1, 27) = 10.73, $p$ < .005), where the average categorisation latency was slower, by 89 msec, in the match than in the mismatch condition.

These results show that the significant interactions were driven by the reversed match effect in the male-pronoun condition, with faster picture categorisation times for mismatching versus matching pictures. In all other conditions, the picture categorisation times were faster for matching versus mismatching sentence-picture pairs. The reversed effect for male faces in the pronoun condition was unexpected. It was not replicated in any the further experiments and remains uninterpretable.

The mixed-model analyses by items revealed no significant main effects for the match factor ($F_2$(1, 20) =0.01, $p$ = .93), the picture factor ($F_2$(1, 20) =1.44, $p$ = .24) and the sentence factor ($F_2$(1, 20) =2.31, $p$ = .15). Further, the interactions were not significant between the sentence and picture factors ($F_2$(1, 20) =0.35, $p$ = .56), between the match and sentence factors ($F_2$(1, 20) =1.51, $p$ = .23), between the match and picture factors ($F_2$(1, 20) =1.68, $p$ = .21) and between the match, sentence and picture factors ($F_2$(1, 20) =2.40, $p$ = .14).

Table 32: Mean picture categorisation times (and standard deviations) in milliseconds per condition

| | Male faces | | Female faces | | Total |
|---|---|---|---|---|---|
| | Pronoun | No-pronoun | Pronoun | No-pronoun | |
| Match | 851 (186) | 755 (181) | 749 (166) | 742 (150) | 773 |
| Mismatch | 762 (144) | 781 (154) | 794 (191) | 765 (138) | 776 |
| Difference (mismatch – match) | -89 | 26 | 45 | 23 | 3 |

## 11.5 Discussion

In this experiment, the only main effect to emerge in the analyses by participants was the effect of the sentence factor, with participants being slower to respond when the pronoun was present versus absent. Note however, that this effect was not replicated in any of the further experiments (and disappeared when the four longest latencies (above 1500 msec) were excluded from the analyses). It was further not confirmed by the item analyses.

The absence of a main effect for the match factor could be due to the reaction times being longer and more variable than usually observed in gender categorisation studies (e.g., Le Gal & Bruce, 2002; Quinn & Macrae, 2005). Participants controlled their own pace of stimulus presentation. They might have intentionally slowed their responding to facilitate task separation because they needed to switch between the different tasks (responding to questions, categorising pictures). If this resulted in participants also separating the reading and picture categorisation tasks, any potential priming effects would have been lost. In light of this potential methodological issue, the results of Experiment 4 cannot be used to draw conclusions as to whether the representations constructed during reading about a stereotype-relevant agent influence subsequent face categorisation. I therefore conducted an additional experiment using a slightly different procedure.

# 12. Experiment 5

## 12.1 Overview

In comparison to Experiment 4, a number of changes were made in the design of this experiment. First, the picture presentation time was controlled by the experimental program (800 msec duration) to encourage (but not require) participants to respond within this time window, thereby decreasing their response latencies. Second, to minimise task switching costs, all trials included the same sequence of tasks and responses: participants first read a sentence, then gender-categorised a face, and finally responded to a comprehension question about the sentence. Third, to minimise between-task response-key confusions between the reading and the comprehension task, the program was adjusted to require a spacebar press for the participant to see the next sentence. Fourth, participants were familiarised with the target faces before the experimental trials to minimise any influence of non-gender-relevant stimulus characteristics. Fifth, a fixation cross at the leftmost position of the screen was included to orient participants to the beginning of each sentence. Finally, a separate practice block of four trials was administered in order to familiarise participants better with the task and to encourage them to increase the pace of their responding in the experimental trials.

## 12.2 Hypotheses

As in Experiment 4, I expected the face categorisation latencies to be faster for experimental sentences containing a sentential subject that matched rather than mismatched the face in gender. In addition, I expected an interaction between the match and pronoun conditions with a stronger priming effect in the pronoun than in the no-pronoun condition. I did not expect any differences in the response latencies to female or male pictures.

## 12.3 Method

### Participants

Twenty participants completed the experiment (5 participants per condition; 13 female; mean age 18.85 years, ranging 18 to 20 years). They were randomly assigned to the four versions of the materials in equal numbers. All participants were undergraduate or postgraduate students at the University of Birmingham with normal or corrected-to-normal vision and native speakers of British English. They received course credits or money in exchange for their participation.

### Apparatus

The apparatus was the same as specified in Experiment 4.

### Materials

The sentence and picture stimuli were the same as in Experiment 4, but whereas in Experiment 4 only half the filler sentences and none of the target sentences were followed by questions, here all 72 sentences were followed with comprehension questions.

### Design

As in Experiment 4, the design was a 2 (picture: female, male) x 2 (sentence: pronoun, no-pronoun) x 2 (match: match, mismatch) within-participants design. The main part of the experiment was preceded by two additional picture-categorisation-only blocks. These were included in order to familiarise the participants with the pictures before the main task and consisted of the 24 experimental and 24 filler pictures in a fixed, random order. The picture-categorisation-only blocks started with four practice trials each. The main part of the experiment was preceded by a short practice block of four trials.

In the main part of the experiment, the main design difference to Experiment 4 was that each of the 24 experimental and 48 filler trials now included the presentation of a sentence, followed by a picture and finally a comprehension question.

**Procedure**

At the beginning of the experiment, participants read about their tasks (see Appendix 33 for instructions). They learned that they would start with the two picture-categorisation-only blocks, followed by three experimental blocks. In these blocks, each trial started with a fixation cross appearing on the left of the screen for 1500 msec, followed by a sentence, then a picture and finally a comprehension question. The sentence-reading and picture-categorisation tasks were self-paced. The pictures disappeared from the screen after 800 msec, encouraging participants to react within this time window. The comprehension question appeared after the participant had responded to the picture.

For the responses to the pictures and comprehension questions, the same keys were specified as in Experiment 4. For moving on to the next sentence, participants pressed the spacebar. The experiment lasted approximately 20 minutes. After completion, the experimenter thanked and debriefed the participants.

## 12.4 Results

### Comprehension question results

The 1440 comprehension questions were responded to correctly in 91% of the cases (1309). Responses were correct for 437 of 480 experimental trials and 872 of 960 filler trials. One-sample *t*-tests showed that the number of correct responses was significantly above chance for both experimental trials (*t*(19) = 28.008, *p* < .001) and filler trials (*t*(19) = 41.49, *p* < .001).

### Picture categorisation results

In the categorisation-only practice blocks 1 and 2, 1855 of 1920 pictures were categorised correctly (97%). Responses were correct for 933 of 960 pictures that would appear in the experimental trials of the main part of the experiment and 922 of 960 pictures that would appear in the filler trials.

In the main blocks of the experiment, 1400 of 1440 pictures were categorised correctly (97%). Responses were correct for 467 of 480 experimental trials and 933 of 960 filler trials. One-sample *t*-tests showed that the number of correct responses was significantly above chance for both experimental trials (*t*(19) = 58.00, *p* < .001) and filler trials (*t*(19) = 64.71, *p* < .001). These results show that picture categorisation accuracy in all blocks was very good.

I analysed the picture categorisation times in Blocks 1and 2 to determine whether the mere categorisation times in this experiment were similar to those reported in the face processing literature. Outlying reaction times over 1200 ms were excluded (0.4% of the correct trials)[31]. In Block 1, participants' mean reaction time was 476 msec (SD =

[31] An inspection of the picture categorisation data in the practice blocks of Experiments 5 to 7 revealed that any extreme values were excluded when applying the common outlier criterion of 1200 msec.

111 msec); in Block 2, it was 481 msec (SD = 130 msec). These results are comparable with others found in face sex categorisation tasks (e.g., Quinn & Macrae, 2005, Le Gal & Bruce, 2002).

My hypotheses for the experimental trials in the main part of the experiment (Blocks 3 to 5) were that the categorisation latencies would be faster on trials where the picture matched rather than mismatched the sentential subject in gender. I further expected an interaction between the match and pronoun conditions with a stronger priming effect in the pronoun than in the no-pronoun condition.

The mean categorisation latencies for female and male pictures following a gender-matching or -mismatching sentence in the pronoun and no-pronoun conditions can be found in Table 33. The data contained one extreme outlier value (3025 msec, range 359 msec to 1466 msec for the remaining data) which was excluded from the analyses (0.1% of the correct trials)[32].

A table with the complete ANOVA results for Experiments 4 to 7 can be found in Appendix 36. A 2 (picture: female, male) x 2 (sentence: pronoun, no-pronoun) x 2 (match: match, mismatch) within-participants ANOVA revealed that there was no difference between the latencies in the match versus mismatch conditions ($F_1(1, 19) = 0.24$, $p = .63$). It further showed that there was no difference between the pronoun and no-pronoun conditions ($F_1(1, 19) = 0.86$, $p = .36$) or between female and male pictures ($F_1(1, 19) = 1.50$, $p = .24$). The interaction between match and sentence, however, was marginally significant ($F_1(1, 19) = 4.02$, $p = .060$). This interaction arose because there was a match effect of 38 msec for the pronoun condition and an

[32] An inspection of the picture categorisation data in the experimental blocks of Experiments 5 to 7 revealed that the datasets of each experiment contained extreme outlier values. These were, however, too different for a common outlier criterion. Therefore, for each experiment, extreme outlier values were excluded based on the respective dataset.

effect of 26 msec in the opposite direction for the no-pronoun condition. Separate analyses for the pronoun and no-pronoun conditions, collapsed across the picture factor, revealed no significant difference between the match and mismatch conditions ($F_1$(1, 19) = 3.55, $p$ = .075 for the pronoun condition; $F_1$(1, 19) = 1.43, $p$ = .25 for the no-pronoun condition). None of the other interactions were significant (Match x Picture: $F_1$(1, 19) = 0.20, $p$ = .66; Picture x Sentence: $F_1$(1, 19) = 3.12, $p$ = .094; Match x Sentence x Picture: $F_1$(1, 19) = 0.09, $p$ = .77).

The mixed-model analyses by items revealed no significant main effects for the match factor ($F_2$(1, 20) = 0.00, $p$ = .97), the picture factor ($F_2$(1, 20) = 2.44, $p$ = .13) and the sentence factor ($F_2$(1, 20) = 0.31, $p$ = .59). Further, the interactions were not significant between the sentence and picture factors ($F_2$(1, 20) = 1.03, $p$ = .32), between the match and picture factors ($F_2$(1, 20) = 0.00, $p$ = .99) and between the match, sentence and picture factors ($F_2$(1, 20) = 0.00, $p$ = .99). The only interaction that reached marginal significance was the interaction between match and sentence ($F_2$(1, 20) = 4.02, $p$ = .059). As in the analyses by participants, separate analyses for the pronoun and no-pronoun conditions, collapsed across the picture factor, revealed no significant difference between the match and mismatch conditions ($F_2$(1, 22) = 2.30, $p$ = .14 for the pronoun condition; $F_2$(1, 22) = 1.90, $p$ = .18 for the no-pronoun condition).

Table 33: Mean picture categorisation times (and standard deviations) in milliseconds per condition

| | Male faces | | Female faces | | Total |
|---|---|---|---|---|---|
| | Pronoun | No-pronoun | Pronoun | No-pronoun | |
| Match | 576 (116) | 609 (114) | 570 (118) | 574 (86) | 582 |
| Mismatch | 604 (124) | 583 (118) | 618 (136) | 549 (72) | 589 |
| Difference (mismatch – match) | 28 | -26 | 48 | -25 | 7 |

### 12.5 Discussion

In this experiment, changes were introduced to encourage participants to react faster than in the preceding experiment. This goal was achieved as the overall categorisation latencies decreased from 775 msec in Experiment 4 to 586 msec in this experiment. No differences were observed between picture categorisation latencies for trials in which the picture matched versus mismatched the sentential subject in gender. Further, no differences were observed between the pronoun and no-pronoun conditions or between the female and male picture conditions. As noted, there was an interaction of sentence and pictures that approached significance. For the pronoun condition the expected results occurred: shorter latencies in the match than in the mismatch condition, even though this numeric difference did not reach significance. For the no-pronoun condition, there was an effect in the opposite direction but, again, the effect was not significant. This pattern does not suggest that the sentences systematically affected the face categorisation latencies.

One reason for an absence of a match effect could be that the comprehension questions motivated the participants to read the sentences very carefully and to keep processing the information even after the sentence had disappeared. This might have added an additional cognitive load, possibly masking any sentence-picture matching effects. I therefore conducted another experiment with a very similar setup as in Experiment 5, but without any comprehension questions at the end of the trials.

## 13. Experiment 6

### 13.1 Overview

In Experiment 6, each experimental and filler sentence was followed by a picture of a face, but no comprehension questions were included. The rationale was that the cognitive load of keeping the sentence information in mind during the picture categorisation task might have masked potential differences between the match and mismatch conditions in Experiment 5.

## 13.2 Hypotheses

As in Experiments 4 and 5, I expected the face categorisation latencies to be faster for experimental sentences with a sentential subject that matched versus mismatched the face in gender. I further expected an interaction between the match and pronoun conditions with a stronger priming effect in the pronoun than in the no-pronoun condition. I expected no differences in the response latencies to female or male pictures.

## 13.3 Method

### Participants

Twenty participants completed the experiment (13 female; mean age 19.50 years, ranging 18 to 24 years). They were randomly assigned to the four versions of the materials in equal numbers. All participants were undergraduate or postgraduate students at the University of Birmingham with normal or corrected-to-normal vision and native speakers of British English. They received course credits or money in exchange for their participation.

### Apparatus, Materials, and Design

Apparatus, materials and design were the same as in Experiment 5, except that no comprehension questions were included here.

**Procedure**

The procedure was the same as in Experiment 5, except that the comprehension questions were omitted and the fixation cross was presented for 800 msec, instead of 1500 msec. This was because of the reduced number of tasks, less cognitive effort was required and thus participants were assumed to need less time between trials. See Appendix 34 for participants' instructions.

## 13.4 Results

In the two categorisation-only practice blocks, 1818 of 1920 pictures were categorised correctly (95%). Responses were correct for 917 of 960 pictures that would appear in the experimental trials of the main part of the experiment and 901 of 960 pictures that would appear in the filler trials.

In the main blocks of the experiment, 1397 of 1440 pictures were categorised correctly (97%). Responses were correct for 469 of 480 experimental trials and 928 of 960 filler trials. One-sample *t*-tests showed that the number of correct responses was significantly above chance for both experimental trials ($t(19) = 57.73$, $p < .001$) and filler trials ($t(19) = 76.24$, $p < .001$.

The latencies for correctly gender-categorised pictures in the practice blocks were 514 msec (SD = 126 msec) in Block 1 and 506 ms (SD = 126 msec) in Block 2 (outlying reaction times above 1200 msec (0.5% of the correct trials) were excluded).

My hypothesis for the experimental trials in the main part of the experiment was that the categorisation latencies would be faster for trials in which the picture matched rather than mismatched the sentential subject in gender. I had further expected an interaction between the match and pronoun conditions with a stronger priming effect in the pronoun than in the no-pronoun condition. I did not expect any differences between female and male pictures.

The mean categorisation latencies for the experimental trials can be found in Table 34. The data contained one extreme outlier value (1933 msec; range 267 msec to 1653 msec for the remaining data) which was excluded from the analyses (0.4% of the correct trials).

A table with the complete ANOVA results for Experiments 4 to 7 can be found in Appendix 36. A 2 (picture: female, male) x 2 (sentence: pronoun, no-pronoun) x 2 (match: match, mismatch) ANOVA revealed that there was no difference between the latencies in the match versus mismatch conditions ($F_1(1,19) = 1.20$, $p = .29$). It further showed that there was no difference between the pronoun and no-pronoun conditions ($F_1(1,19) = 1.22$, $p = .28$) or the female and male picture conditions ($F_1(1,19) = 1.00$, $p = .33$). The following interactions were not significant: Match x Sentence ($F_1(1,19) = 0.67$, $p = .43$), Sentence x Picture ($F_1(1,19) = 2.86$, $p = .107$), Match x Sentence x Picture ($F_1(1, 19) = 0.70$, $p = .41$). The interaction between the factors match and picture, however, was significant ($F_1(1, 19) = 5.46$, $p < .05$). Analyses of simple effects showed that for male pictures there was no difference between the match or mismatch conditions ($F_1(1, 19) = 1.03$, $p = .32$). The analysis for female pictures revealed a marginal difference between the match and mismatch conditions (526 msec versus 488 msec; $F_1(1, 19) = 4.10$, $p = .057$). As Table 34 shows, there was a substantial difference between the match and mismatch conditions in the pronoun condition, but a much smaller difference in the no-pronoun condition. Following up this difference, separate analyses for the female pronoun and no-pronoun conditions revealed that the differences between the match and mismatch conditions were not significant in either of these conditions ($F_1$(1, 19) = 2.69, $p$ = .12 for the pronoun condition; $F_1$(1, 19) = 0.55, $p$ = .47 for the no-pronoun condition).

The mixed-model analyses by items revealed no significant main effects for the match factor ($F_2$(1, 20) = 1.08, $p$ = .31), the picture factor ($F_2$(1, 20) = 0.84, $p$ = .37) and the sentence factor ($F_2$(1, 20) = 0.81, $p$ = .38). Further, the interactions were not significant between the sentence and picture factors ($F_2$(1, 20) = 1.34, $p$ = .26), between the match and sentence factors ($F_2$(1, 20) = 0.69, $p$ = .42) and between the match, sentence and picture factors ($F_2$(1, 20) = 0.74, $p$ = .40).

However, the interaction between the match and picture factors was significant ($F_2$(1, 20) = 5.04, $p$ < .05). Separate analyses for male and female pictures across sentence conditions revealed no significant differences between match and mismatch conditions for male pictures ($t$(10) = 0.94, $p$ = .37), but a marginal difference between match and mismatch conditions for female pictures ($t$(10) = 2.14, $p$ = .058). Separate analyses for the female pronoun and no-pronoun conditions revealed that the differences between the match and mismatch conditions were not significant in either of these conditions ($t$(10) = 2.02, $p$ = .071 for the pronoun condition; ($t$(10) = 0.53, $p$ = .61 for the no-pronoun condition).

Table 34: Mean picture categorisation times (and standard deviations) in milliseconds per condition

| | Male faces | | Female faces | | Total |
|---|---|---|---|---|---|
| | Pronoun | No-pronoun | Pronoun | No-pronoun | |
| Match | 488 (71) | 490 (79) | 552 (164) | 499 (88) | 507 |
| Mismatch | 499 (63) | 503 (104) | 491 (77) | 484 (73) | 494 |
| Difference (mismatch – match) | 11 | 13 | -61 | -15 | -13 |

## 13.5 Discussion

In Experiment 6, no significant differences were observed in the response latencies to pictures in the match versus the mismatch or the pronoun versus no-pronoun conditions. The significant interaction between the match and picture factors arose because within the male picture condition, the categorisation latencies were shorter in the match than in the mismatch condition (489 msec versus 501 msec), whereas within the female picture condition, the latencies were longer in the match than in the mismatch condition (537 msec versus 488 msec). A follow-up post-hoc analysis revealed, however, that neither of these differences was significant.

The reaction times in this experiment were overall shorter than in Experiment 5. This indicates that switching between categorising faces and responding to comprehension questions had indeed slowed participants' categorisation responses. The null results in this experiment, however, suggest that it was not this slowdown that accounted for the absence of differences between the match and mismatch conditions in Experiment 5.

Experiments 4 to 6 tested whether the processing of stereotype-relevant information during sentence reading would exert an influence on the gender categorisation of faces. The experiments yielded no effect of sentence and picture gender match, suggesting that there was no cross-modal priming from sentences to face categorisation.

The sentence stimuli had been presented in two versions. One version included a gender stereotype-relevant occupation label; the other version included, in addition, a reflexive pronoun toward the end of the sentence. This was to test whether potential facilitation effects by stereotype-relevant occupation labels would be enhanced by additional activation of stereotype-matching gender features through the presence of a reflexive pronoun. Because no differences were observed in any of the experiments between gender-matching and -mismatching sentential subjects and pictures, regardless of whether or not a reflexive pronoun was present, this question remains unresolved.

One reason for a missing effect of the stereotypical gender of the sentential subject on face categorisation latencies could be that linguistic stereotype processing does not have a generalising cross-modal effect on non-linguistic processes. This explanation, however, would contradict earlier findings of such effect (e.g., Kawakami & Dovidio, 2001; Lemm et al., 2005). The reason could also be that the materials used in these experiments were too complex. Apart from the stereotype-relevant information provided by the occupation labels and pronouns, the sentences included other information, some of which followed the labels and pronouns. The processing of this information might have masked the influence of the stereotype-relevant information on the picture categorisation latencies. I therefore tested in Experiment 7 whether reading stereotype-relevant information could yield cross-modal effects on picture categorisation times when simpler stimuli were used as primes. For this, I used the bare occupation labels.

# 14. Experiment 7: Effects of stereotype activation during word reading on face processing

## 14.1 Overview

In Experiment 7, participants read words signifying gender-stereotypical occupation labels (e.g., “babysitter”, “pilot”) that were immediately followed by female or male faces. Their task was to categorise the faces by gender.

## 14.2 Hypotheses

In previous studies examining cross modal effects from linguistic to pictorial information, the stimuli were not sentences but words. Lemm et al. (2005) used gender-stereotype-relevant role words (e.g., “mechanic”, “hairdresser”) as primes, and Kawakami and Dovidio (2001) used gender-stereotypic traits (e.g., “caring”, “technical”). These studies showed that with such simple kind of stimuli, cross-modal effects from linguistic information to gender-relevant pictures were possible. Based on these findings, I expected the face categorisation latencies to be faster in trials in which the stereotypical gender of the occupation label matched rather than mismatched the face.

## 14.3 Method

### Participants

Twelve participants completed the experiment (7 female; mean age 19.66 years, ranging 18 to 21 years). They were randomly assigned to the two versions of the materials in equal numbers. All participants were undergraduate or postgraduate students at the University of Birmingham with normal or corrected-to-normal vision and native speakers of British English. They received course credits or money in exchange for their participation.

### Apparatus

The apparatus was the same as specified in Experiment 4.

### Materials

The picture stimuli were the same as in Experiments 4. The written stimuli consisted of 24 bare nouns, signifying stereotypically female or male occupations. The occupation labels were the same as in the experimental sentences in Experiment 4. The 48 written filler stimuli consisted of concrete nouns (e.g., stamp, dog, clothes). None of the filler words referred to people.

### Design

The design was based on Experiments 6. Three blocks in the main part of the experiment were preceded by two picture-familiarisation blocks. In the picture-familiarisation blocks, the 24 experimental and 24 filler pictures were presented in a fixed random order. In the trials of the main part of the experiment, the 24 experimental sentences of Experiments 4 to 6 were replaced with the stereotype-relevant occupation names, and the 48 filler sentences with 48 filler words. In half of the experimental trials, the stereotype-relevant occupation and the face matched in gender; in the remaining trials, they mismatched in gender. This resulted in a 2 (picture: female; male) x 2 (match: match, mismatch) within-participants design.

### Procedure

The procedure was the same as in Experiment 6, only with different written stimuli (see Appendix 35 for participants' instructions). The trials consisted of a word and a picture, preceded by a fixation cross, which was presented for 800 msec. The word reading time was fixed to 200 msec. The picture categorisation task was self-paced; however, the pictures disappeared from the screen after 800 msec. The experiment lasted about 15 minutes. After completion, participants were thanked and debriefed.

## 14.4 Results

In practice blocks 1 and 2, 1110 of 1152 pictures were categorised correctly (96%). Responses were correct for 559 of 576 pictures that would appear in the experimental trials of the main part of the experiment and 561 of 576 pictures that would appear in the filler trials.

In the main blocks of the experiment (Blocks 3 to 5), 833 of 864 pictures were categorised correctly (96%). Responses were correct for 280 of 288 experimental trials and 553 of 576 filler trials. One-sample *t*-tests showed that the number of correct responses was significantly above chance for both experimental trials ($t(11) = 34.00$, $p < .001$) and filler trials ($t(11) = 55.48$, $p < .001$).

For the picture-categorisation analysis for Blocks 1 and 2, reaction times above 1200 msec were excluded (0.5% of the correct trials). In Block 1, the mean reaction time for correct responses was 508 msec (SD = 117 msec). In Block 2, the mean reaction time for correct responses was 498 msec (SD = 108 msec).

My hypothesis for the experimental trials in the main part of the experiment (Blocks 3 to 5) was that the categorisation latencies would be shorter on trials in which the picture matched rather than mismatched the stereotypical gender of the occupation role. The mean categorisation latencies for female and male pictures following a gender-matching or -mismatching occupation label can be found in Table 35. The data contained one extreme outlier value (1266 msec; range 360 msec to 986 msec for the remaining data) which was excluded from the analyses (0.1% of the correct trials). A table with the complete ANOVA results for Experiments 4 to 7 can be found in Appendix 36. A 2 (picture: female; male) x 2 (match: match, mismatch) within-participants ANOVA revealed a significant difference between the latencies in the match or mismatch conditions ($F_1(1, 11) = 19.66$, $p < .005$), such that participants were faster to categorise faces in the match than mismatch condition. No difference was found between the female and male picture conditions ($F_1(1, 11) = 0.35$, $p = .57$)

and there was no interaction between the match and the picture factors ($F_1$(1, 11) = 1.48, $p$ = .25).

The between-items analyses by items revealed a significant main effects for the match factor ($F_2$(1, 20) = 20.95, $p$ < .001). No difference was found between the female and male picture conditions ($F_2$(1, 20) = 0.04, $p$ = .84). There was no significant interaction between the match and picture factors ($F_2$(1, 20) = 1.04, $p$ = .319).

Table 35: Mean picture categorisation times (and standard deviations) in milliseconds per condition

| | Male face | Female face | Total |
|---|---|---|---|
| Match | 511 (72) | 503 (74) | 507 |
| Mismatch | 556 (67) | 577 (98) | 567 |
| Difference<br>(mismatch – match) | 45 | 74 | 60 |

### 14.5 Discussion

Experiment 7 tested whether reading stereotype-relevant occupation labels can exert an influence on the processing of female and male faces. The results showed that this is the case: Participants were faster to indicate the gender of a face when it matched the stereotypical gender of the preceding occupation label than when it mismatched the stereotypical gender of the preceding occupation label. For example, participants were faster to correctly categorise a female face when preceded by the word *secretary* than by the word *mechanic*. It appears that the absence of gender matching effects in the preceding experiments was in some way linked to the complexity of the linguistic primes.

# 15. General discussion

## Summary of the results

In Chapter 3, I found that the episodic representations constructed during reading of stereotype-relevant information affected further linguistic processing. In this chapter, I sought to investigate whether the scope of these representations as interpretative frameworks for new incoming information was restricted to linguistic information or whether it extended to non-linguistic processing. I did this by combining a reading tasks including stereotype-relevant information with a non-linguistic face-processing task. The stimuli in the reading task were sentences in Experiments 4 to 6 and bare noun words in Experiment 7. In the experiments including sentence stimuli, no consistent cross-modal facilitation effect from the reading stimuli to the picture categorisation times was observed. In Experiment 7, however, such an effect emerged, suggesting that cross-modal facilitation from stereotype-relevant words to pictures is possible if the stimuli are simple. This means that reading about stereotype-relevant information can have an influence on non-linguistic cognitive processes.

## Placing the findings within the context of prior research

Similar cross-modal priming effects from stereotype-relevant words to pictures as observed in Experiment 7 have been found in previous studies. Lemm et al. (2005) used gender-stereotype-relevant occupation words (e.g., “mechanic”, “hairdresser”) as primes and line drawings of women and men as targets. They found a significant interaction of prime and target gender: For female targets, responses were facilitated in response to gender-matching versus -mismatching and -neutral primes; for male targets, responses were slower in response to gender-mismatching versus -matching and -neutral primes. Experiment 7 extends these findings in that photographs were used instead of line-drawings and a solid priming effect was found which did not interact with the gender of the face. Kawakami and Dovidio (2001), too, found a cross-modal priming effect using gender-stereotypic traits as prime words and photographs of female and male faces as targets to assess the reliability of implicit stereotyping. They found that female stereotypes facilitated the categorisation of female faces and that male stereotypes facilitated the categorisation of male faces. The results of these studies and Experiment 7 show that reading about stereotype-relevant information can influence not only the processing of further linguistic, but also non-linguistic, pictorial information. What had not been addressed before was the scope of this cross-modal influence. The results of the experiments in this chapter showed that whereas a cross-modal priming effect could be observed with bare nouns, no such effect emerged with sentence stimuli.

**Possible explanations for the absence of a sentence effect**

With hindsight, Experiment 6, with its methodological improvements over Experiments 4 and 5, seemed the most likely of the three experiments to yield sentence-picture-matching effects, yet none were found. One possible explanation for this could be that the removal of the comprehension questions in this experiment led participants to read the sentences less carefully, possibly leading them to overlook the crucial stereotype-relevant information. To investigate this possibility, I compared the mean sentence reading times (time interval between presentation onset and button press) in Experiments 4 to 6, which are assumed to reflect participants' processing time and effort. Outlying reading times below 1000 ms or above 6500 ms were excluded from the analyses, which resulted in a data loss of 1% for each experiment. The mean reading times were 2982 msec (SD = 1004 msec, N = 667) in Experiment 4, 2847 msec (SD = 898 msec, N = 475) in Experiment 5, and 2570 msec (SD = 865 msec, N = 476) in Experiment 6. It can be seen that the reading times in Experiment 6 were slightly shorter than in the other experiments. However, this experiment also included fewer tasks and therefore lower task-switching demands. The average reading time per word of 239 msec (2570 msec divided by an average of 10.75 words per sentence) is in a range that suggests processing to a level of comprehension. It is therefore unlikely that the absence of a match effect between sentential subject and picture gender was due to participants not reading the sentences properly.

Another reason why no cross-modal sentence-picture match effect was observed could be that the sentences contained more and more complex information than the bare nouns. For example, a sentence like “On several occasions the receptionist hurt herself with the sharp scissors“ activates not only the concept for “receptionist”, but also the concepts for “several occasions”, “hurt”, “sharp” and “scissors”. The activation of these additional concepts might have masked any gender-priming effects. This seems plausible given that the nouns signifying the agents were always positioned near the beginning of the sentences, followed by the verb and several other words.

Also, it was possible that higher-level inference processes about information other than the stereotype-relevant information might have masked possible picture-facilitation effects. It has been shown that reading about an actor displaying a single behaviour can prompt spontaneous trait inferences by the observer (for a review, see Uleman, Newman, & Moskowitz, 1996). Most of the sentences describe a single event or action. Participants might, for example, have engaged in processes like inferring the trait “clumsy” when reading about the receptionist. Such processes might have interfered with any effects of the stereotype-relevant information on the picture categorisation times.

In sum, it seems that when other concepts are activated (directly via reading the sentence or indirectly via the inferences drawn from the sentence), stereotypes may not be the only determinant of how subsequent information is processed. One might say that the participants created episodic representations of the events described in the sentences, but that in these representations the gender of the agent was not very salient. Interestingly, this was true even when the gender was highlighted by the presence of a reflexive pronoun.

**Interpretation within the working model versus simulation model**

There are different explanations of how reading information could influence face categorisation. The explanation within the framework of the working model is that picture categorisation times are facilitated due to a match of the conceptual gender features of the occupation label and the picture. When the occupation word (e.g.., *mechanic*) is read, the prototype representation is activated. This representation has a link to the stereotype-matching gender feature node (*male*). When the picture, for example of a man, is encountered, the prototypical conceptual representation of a man with its associated *male* gender node is activated. This gender node will be part of the newly constructed episodic representation of the picture. In case the occupation label and the picture match in gender, the episodic picture representation also shares the pre-activated agent gender feature. This pre-activation facilitates the construction of an episodic representation of the picture and speeds up gender categorisation times. Another explanation of how reading could influence face categorisation is that upon encountering the occupation label, pictorial features of the job holder are simulated which then facilitates the processing of the face. This approach is informed by the situated simulation theory (Barsalou, 2008), according to which conceptual representations are multi-modal simulations. According to this approach, when presented with gender-stereotype-relevant occupation labels, participants construct a representation that includes a pictorial simulation of gender-typical facial features. When then seeing a face with features that overlap with the simulation, gender categorisation will be facilitated in comparison to when the facial features do not overlap with the simulation. The results of Experiment 7 are consistent with both the working model and the simulation approach.

### Social relevance of the findings

The findings in this chapter are socially relevant because they show that reading simple stereotype-relevant information can subsequently influence the perception and categorisation of members of the stereotyped group. They also show, however, that these effects are limited, as they are attenuated when the linguistic information is more complex. It would be worth considering this finding for the design of other studies investigating the effects of linguistic stimuli on picture processing—especially social psychological studies, where a dominant theme is that stereotypes automatically and inevitably shape processing (Bargh, 1999).

# Chapter 5
# General Discussion

The topic of this thesis is the processing of stereotype-relevant information during reading. In particular, I have been interested in the resource-dependency of stereotype-mismatch detection and resolution. In addition, I investigated the effects of stereotype-relevant episodic representations on subsequent linguistic and non-linguistic processing and memory.

In Experiment 1, I replicated the mismatch effect found in previous studies (e.g., Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Sturt, 2003). I also tested the effects of a concurrent 5-digit retention task on online processing of and memory for stereotype-relevant information. In Experiments 2 and 3, I investigated the role of the episodic representation constructed during reading stereotype-relevant information as interpretative framework for the processing of further linguistic information and for memory. For this, I added a further sentence with stereotype-relevant information to the stereotype-relevant sentences of Experiment 1. I tested whether after reading the first sentence, a mismatch effect would still occur in the second sentence. In Experiments 4 to 7, I examined the role of the stereotype-relevant episodic representation as interpretative framework on further non-linguistic processing. Here, I measured gender-categorisation times for pictures of faces that matched or mismatched the stereotypical gender of the occupation holder in the preceding linguistic context. In the following section, I summarise and discuss my findings. I then outline how the assumptions of my working model can be formulated in connectionist terms. Finally, I highlight the broader implications of my findings.

# 16. Summary and discussion of the findings

## 16.1 Overview of the findings

An overview of the main results can be found in Table 36.

Table 36: Overview of the main results

| Experiment | Example | Online effects | Memory effects |
|---|---|---|---|
| Experiment 1,<br>no-load condition | “Last week the *secretary* familiarised *herself/ himself* with the new photocopier” | - Early and late mismatch effects,<br>- Late load effects<br>- No Match x Load interaction | - Above-chance memory sensitivity,<br>- Conservative response bias towards stereotype-matching information<br>- Greater response bias under cognitive load |
| Experiment 1,<br>load condition | [Load number, e.g., *51278*]<br>“Last week the *secretary* familiarised *herself/ himself* with the new photocopier”<br>[Probe number, e.g., *51298*] | | |
| Experiment 2,<br>token condition | “The *elderly secretary* thoroughly familiarised *herself/ himself* with the new computer a few months before retiring. To everyone’s surprise, the *secretary* really enjoyed *herself/ himself* while exploring the potential of the computer.” | - Early and late mismatch effects,<br>- Early and late sentence effects<br>- No Match x Sentence interaction<br>- No token-type effects | n.a. |
| Experiment 2,<br>type condition | “The *elderly secretary* reluctantly familiarised *herself/ himself* with the new computer a few months before retiring. In contrast, the *new secretary* really enjoyed *herself/ himself* while exploring the potential of the computer.” | | |
| Experiment 3 | “The elderly *secretary* thoroughly familiarised *herself/ himself* with the new computer a few months before retiring and, to everyone’s surprise, really enjoyed *herself/ himself* while exploring the potential of the computer.” | - Early and late mismatch effects<br>- Early and late effects of pronoun number<br>- Interactions between match and pronoun number on later measures | - Above-chance memory sensitivity,<br>- Conservative response bias towards stereotype-matching information<br>- Greater response bias for matching than mismatching sentences |
| Experiments 4 to 6, pronoun condition | “Last week the *secretary* familiarised *herself* with the new photocopier.”<br>+ female/ male face | - No effects for gender-match between sentence and picture<br>- No pronoun effects | n.a. |
| Experiments 4 to 6, no-pronoun condition | “Last week the *secretary* became familiarised with the new photocopier.”<br>+ female/ male face | | |
| Experiment 7 | “secretary”<br>+ female/ male face | - Effects for gender-match between occupation label and picture | n.a. |

## 16.2 Resource-dependency of mismatch detection

From previous studies (e.g., Duffy & Keir, 2004; Kreiner, Sturt, & Garrod, 2008; Sturt, 2003), it has not been clear, whether the detection and resolution of stereotype-mismatching information during reading is resource-dependent or whether it can still take place when readers are under cognitive load. In Experiment 1, I found that the online mismatch effect was unaffected by the cognitive load of a 5-digit retention task. This result was not caused by an overall absence of influence of the cognitive load manipulation on online processing, as can be seen in the main effect on total reading time. It can therefore be concluded that this kind of cognitive load does not affect readers' ability to detect stereotype-mismatching information.

It is difficult, however, to assess the wider implications of these findings for the automaticity of stereotype-mismatch detection during reading in general, as the effect of the cognitive load manipulation on reading time was contrary to my expectations. Belke (2008) found that the 5-digit retention task slowed word-naming latencies. Based on these results, I had expected *longer* overall sentence reading times due to the additional cognitive demands. However, the imposition of a cognitive load caused the participants in Experiment 2 to read *faster* than in the no-load condition, presumably in order minimise the time they had to maintain the number. In this respect, the load manipulation induced time pressure on the online processing. The effects of other cognitive load manipulations that slow down but do not otherwise impair the reading process on mismatch detection and resolution remain to be tested.

## 16.3 Effects of stereotype-relevant episodic representations on subsequent processing

### Effects on subsequent linguistic processing

In Experiments 2 and 3, I investigated the strength and stability of the representations constructed during reading stereotype-relevant information by examining their effect on subsequent linguistic processing. In Experiment 2, I appended a second sentence that repeated the stereotype-relevant information. Contrary to my expectations, a mismatch effect occurred not only in the first but also in the second sentence. The effect in the second sentence could have arisen because the episodic representations constructed during reading the first sentence were not strong or stable enough to override the gender-stereotypical representations. Alternatively, it could have arisen because the occupation label was repeated at the beginning of the second sentence, which might have reactivated the stereotype. To disambiguate the source of the second mismatch effect, the reference to the agent was left implicit in Experiment 3. The results of this experiment showed an interaction between match and pronoun number. Planned comparisons revealed that the interaction was due to a mismatch effect on the first pronoun, but an absence of such an effect on the second pronoun. These results suggest that the repeated mismatch effect in Experiment 2 had been due to the reactivation of the stereotype by the second mention of the occupation label. However, the repetition of the occupation label was not the only difference between Experiment 2 and 3: The stimuli in Experiment 2 consisted of two sentences, whereas the stimuli in Experiment 3 consisted of only one sentence, which made it possible to keep the reference to the agent implicit. Participants in Experiment 2 might have treated each sentence as an entity. This might have contributed to the repeated mismatch effect in the second sentence. Unfortunately, it is not easy to examine such an effect of processing strategy within the context of these experiments because the syntactic structure of sentences with an explicit or implicit reference back to an initially introduced agent necessarily differ. However, it is more likely that the mismatch effect observed for the second sentence in Experiment 2 was due to the repetition of the occupation label rather than the use of two separate sentences, as it has previously been shown that the resolution of expectancy-violating information in one sentence can have an effect on the processing of expectancy-violating information in subsequent sentences. Nieuwland and Van Berkum (2006), for example, found that orally presented expectancy-violating information embedded in a sentence such as ‘‘Once upon a time, a psychotherapist was consulted in her home office by a *yacht* with emotional problems”, elicited a N400 effect[33] which is argued to reflect a response to semantic anomalies (van Berkum, Hagoort, & Brown, 1999). However, if the expectancy-violating information was presented later in the discourse context once participants had made sense of the locally expectancy-violating information—for example in a subsequent sentence like “At that moment the *yacht* cried out that he was absolutely terrified of water”—no N400 effect was found. This finding argues against the view that sentences within the same discourse context are first treated as separate entities and only later semantically integrated. It speaks instead for the incremental integration of information within the same discourse context, even across sentences. It is therefore likely that the differential effects in Experiments 2 and 3 were due to the repetition of the occupation label in Experiment 2, rather than the syntactic differences, especially considering that in both experiments the stimuli were

[33] [33] The N400 is a negative wave with an onset at about 200 msec after the onset of a critical word and a peak at about 400 msec after the onset of the word (van Berkum, Hagoort, & Brown, 1999).

presented on one screen and therefore within one processing episode, encouraging integration.

The results of Experiment 3 confirm previous results showing that the episodic representations constructed during reading prior disambiguating context can result in the cancellation of a mismatch effect (e.g., Carreiras, Garnham, Oakhill, & Cain, 1996; Duffy and Keir, 2004; Kreiner, Sturt, & Garrod, 2008). They also show that such a cancellation is possible not only with very explicit, but also rather subtle disambiguation information, such as a reflexive pronoun. A comparison of the results of Experiments 2 and 3 and with previous research indicated, however, that the episodic representations constructed from subtle disambiguating information are more susceptible to being overridden by stereotypical representations than episodic representations constructed from more salient disambiguating information (see chapter 3, section 9.1). This could be because participants might be less likely to integrate the disambiguating information and construct an accurate episodic representation when the disambiguating information is subtle than when it is more explicit. Additionally, even when accurate representations have been constructed, the likelihood that some participants forget them on some of the trials before they encounter the next piece of stereotype-relevant information is higher when the disambiguating information is subtle than when it is more explicit.

### Effects on non-linguistic processing

After Experiment 3 had shown that the episodic representations constructed during reading stereotype-relevant information can influence further linguistic processing, Experiments 4 to 7 tested whether this influence could extend to non-linguistic information. In Experiments 4 to 6, no differences were found between the gender-categorisation latencies for faces that matched versus mismatched the stereotypical gender of the agent in the preceding sentence. The results of these experiments seemed to suggest that the episodic representations constructed during reading stereotype-relevant sentences do not exert an influence beyond further linguistic processing to face processing. However, findings of previous studies (e.g., Kawakami & Dovidio, 2001; Lemm, Dabady, & Banaji, 2005) suggested that such cross-modal priming effects are possible with simple word stimuli. In Experiment 7, the participants' task was therefore to read bare stereotype-relevant occupation labels before gender-categorising female and male faces. In this experiment, an effect of label-face match was found. The differential effects of priming following single words but not following sentences could be due to a variety of reasons.

One reason for the lack of priming from sentences to faces might be attentional in nature. When participants processed the sentences, it is likely that the entire sentence representation was their focus of attention. It is possible that in this context, they did not pay enough attention to the occupation label for an effect of priming to occur. When the occupation label was presented on its own, however, it would have been in the focus of attention and anything that could be associated with it would have had increased potential to be influenced. This might have led participants to integrate or compare the pictures with the linguistic information in Experiment 7, but not in Experiments 4 to 6.

Another potential reason for the difference was that in Experiment 7, only the concept associated with the occupation label was activated, whereas in Experiments 4 to 6, additional concepts were triggered. For example, during reading the sentence "On Saturday the cheerleader dressed herself in a bright costume", apart from the concept for "cheerleader", the concepts for "Saturday", "dressed", "bright" and "costume" would be activated. Further, it is likely that within the sentence context participants also engaged in higher-level inference and integration processes (see section 15). For example, previous research has shown that reading about a behavior can prompt observers to make spontaneous trait inferences (for a review, see Uleman, Newman, & Moskowitz, 1996). Participants in Experiments 4 to 6 might have engaged in such higher-level inference processes during sentence reading (e.g., inferring the trait "vain" when reading about the cheerleader). It is possibly that the activation of multiple concepts and higher-level processing during sentence reading resulted in the dilution of the priming effect.

These accounts for the differences between word and sentence priming effects fit with the distinction of functionally different regions within working memory made by Oberauer (2002). According to Oberauer, a "*region of direct access* holds a limited number of chunks available to be used in ongoing cognitive processes" and a "*focus of attention* holds at any time the one chunk that is actually selected as the object of the next cognitive operation" (p. 412; emphasis added). In Experiments 4 to 6, the occupation label was only one part of the information that was kept available for ongoing processing (i.e., it was within the region of direct access) and might therefore not have been focal enough to render a cross-modal face priming effect. In Experiment 7, however, the occupation label was the only focus of attention and was therefore a strong enough prime for any subsequent relevant information.

Another account for the differential effects for word and sentence primes could be that the time interval between the occupation label and the face was longer in Experiments 4 to 6 than in Experiment 7, which might have masked a gender-priming effect. The findings in Experiments 4 to 6 do not to support this interpretation, however: No differences were found between the pronoun and no-pronoun conditions which differed in the proximity of gender information to the face.

In sum, more research is needed to distinguish between the different accounts about why bare occupation labels but not sentence primes facilitated face-categorisation times.

### Discussion of the differential effects on linguistic and non–linguistic processing

In Experiment 3, an influence of the episodic representation constructed during reading stereotype-relevant information on subsequent processing was observed. A mismatch effect emerged on the first but not the second pronoun within the same processing episode. In terms of my working model, I explained this by assuming that the episodic representation formed after the encounter with the first pronoun was strong and stable enough to be maintained and remain active until the encounter with the second pronoun (see section 9.1). In contrast, in Experiments 4 to 6, I did not find an effect of the episodic representation formed during reading a sentence on subsequent processing of female or male faces. It is difficult to isolate a reason for these differential results, as the target stimuli (sentence versus picture) and tasks (reading versus face categorisation) differed in many ways. Further, the design of the experiments was different with participants reading only matching sentences in Experiments 4 to 6, but reading both matching and mismatching sentences in Experiment 3. The latter might have increased the salience of the gender feature, resulting in increased attention allocation to stereotype-relevant information in this task compared to Experiments 4 to 6. Additionally the dependent measures were different: In Experiment 3, participants' online processing, reflected in their eye movements, was investigated, whereas in Experiments 4 to 6, picture categorisation latencies were measured. Another difference between the experiments was that qualitatively different effects were expected: a cancellation of an online mismatch effect during reading in Experiment 3 versus an occupation label-face gender-match effect in Experiments 4 to 6.

Despite the difficulty to pinpoint a single reason for the differential effects of sentence stimuli on subsequent linguistic and non-linguistic processing, one plausible account is that participants might have employed different processing strategies. In Experiment 3, the critical second pronoun was part of the same text as the introducing stereotype-relevant information (occupation label and first pronoun). Therefore, the priming information and the target were part of the same task. This might have motivated participants to keep the episodic representation constructed during the first reading of agent and pronoun actively in working memory and to integrate any further information. In Experiments 4 to 6, however, the sentence reading and picture-categorisation tasks were very likely perceived as separate processing events. As I did not want participants to realise the purpose of the experiments, I encouraged this perception with the instructions (see Appendices V – X). Participants might therefore not have been motivated to keep the stereotype-relevant episodic representation actively in working memory and compare it to or integrate it with the pictorial information in the face-categorisation task.

Another account for the differential findings in Experiments 4 to 6 and Experiment 3 is the difference in modalities of the subsequent information in the experiments. Lemm and colleagues put forward the claim that "cross-modality priming requires stronger underlying prime–target relationships to produce a priming effect compared with same-modality priming" (p. 223, see also Federmeier & Kutas, 2001). The differences between same- and mixed-modality priming effects could be due to the fact that words and pictures are processed in different semantic systems (e.g., Paivio, 1971, 1986) and consequently the priming effects within one modality would be stronger than priming effects across modalities. However, as I have discussed in chapter 4 (see section 10), I assume that words and pictures share amodal conceptual representations (see also Bajo, 1988; Federmeier & Kutas, 2001; Kroll, 1990; Potter & Kroll, 1987; Potter, Kroll, Yachzel, Carpenter, & Sherman, 1986). Within this approach, the finding of stronger same- compared to mixed-modality priming effects can be explained, for example, by the benefits of similarities within same-modality prime-target pairs that are not conceptual in nature, such as visual and lexical similarity (Federmeier & Kutas, 2001). In my experiments, the same-modality priming effect could have been stronger because the word on which the cancellation effect manifested in Experiment 3 (*herself* or *himself*) was a repetition of a word that had already been processed and that had contributed to the episodic representation of the agent. It was therefore already pre-activated on a visual, lexical and conceptual level, which might have made the reactivation of the word and the associated episodic representation easier. The facilitated reactivation of the episodic representation might have contributed to it outweighing the semantic prototype representation and therefore contributed to the cancellation of a second mismatch effect. A face in Experiments 4 to 6, however, was an entirely new item that could not benefit from such a reactivation.

One way of evaluating whether the differential effects on subsequent processing in Experiments 3 versus 4 to 6 were due to the processing strategy account (i.e., participants' ongoing processing and integration effort for the second pronoun, but not the face) or the cross-modality account (i.e., reduced cross-modality compared to same-modality facilitation) would be to present sentence and picture stimuli within rebus sentences. This could be instantiated by introducing an agent with an occupation label and pronoun at the beginning of a sentence and later referring to the agent with a picture of a female or male face. Participants' task could again be to gender-categorise the picture. If the null effect in Experiments 4 to 6 was due to reduced cross-modality facilitation, one would expect that categorisation latencies would not differ for pictures that matched versus mismatched the pronoun. If the null effect in Experiments 4 to 6 was due to the picture not being part of an ongoing integration process within the same task, however, one would expect that the presentation of the face within the sentence might make it more relevant to the ongoing processing of the text representation, kept active in working memory. Participants might also be more motivated to integrate the picture with the sentence. In this case, a facilitation of categorisation times for pictures that matched rather than mismatched the gender of the sentential agent might be found. Results by Potter and colleagues (1986) that participants in their task understood and remembered rebus sentence almost equally well as regular sentences (see chapter 4, section 10) might give an indication that facilitation effect could be found for picture-categorisation within rebus sentences.

## 16.4 Memory for stereotype-relevant information

Previous studies into the mismatch effect have not assessed whether participants could remember the stereotype-relevant information later, or still access the stereotype-relevant episodic representations constructed during reading. I tested this in Experiments 1 and 3, where I administered memory questionnaires after the reading tasks. In Experiment 1, participants indicated whether a particular agent (e.g., *secretary*) had been female or male. Half of the participants had been under cognitive load during encoding. In Experiment 3, participants decided whether they had already seen a sentence or whether it was new. Old and new sentences differed only in the reflexive pronoun. As an improvement over the questionnaire in Experiment 1, this task enabled the comparison of sensitivity and bias between originally matching and mismatching sentences.

In both experiments, memory sensitivity was above chance for both matching and mismatching sentences. This result indicates that the mismatch effect might not only reflect the detection, but also the resolution, of the stereotype-mismatching information, resulting in accurate episodic representations. It extends the online findings by confirming that these episodic representations can be maintained beyond the context of a processing episode.

In addition, in both Experiments 1 and 3, participants exhibited a conservative response bias. That means that regardless of which item had originally been presented, participants tended to favour stereotype-matching responses. This result indicates that participants consulted their stereotypes when unable to remember the episodic representations and used the semantic stereotype representations as guessing aid. Also, in terms of my working model, the presentation of a stereotype-matching (questionnaire) item triggers the (re)activation of the stereotype-matching prototype representation. Experiment 2 demonstrated that the activation of a prototype representation can challenge the activation of an episodic representation. It could therefore be that the activation of the stereotypical prototype representation by a matching questionnaire item competed with the episodic representation constructed during reading, resulting in the tendency to respond in the stereotype-matching way. It would be worth considering this finding for the design of other questionnaires assessing memory for stereotype-relevant information. For example, instead of presenting participants with stereotype-matching or -mismatching response options, it might be better to ask neutral questions, as was the case in Experiment 1 (e.g., “Was the secretary female/ male?”).  Alternatively, instead of using a recognition questionnaire, free recall could be measured to assess memory for stereotype-relevant information, which makes the use of cues unnecessary.

Experiment 1 showed that the tendency to respond in a stereotype-matching way increased under cognitive load, whereas online processing was unaffected by cognitive load. The dissociation of the effects of cognitive load on memory and online processing in Experiment 1 could indicate that the memory data may reflect on cognitive processes not captured by the reading times, namely late effects of context integration. In section 5.5, I argued that the increased response bias under cognitive load in Experiment 1 might be a result of the processing-by-product principle (Carlston & Smith, 1996, as cited in Smith, 1998). According to this principle, the ease of reconstructing an exemplar representation corresponds to the effort allocated to its formation. I also suggested that cognitive load might affect late integration processes, based on the main effects of cognitive load on late eye-movement measures. If these late integration processes are affected, fewer cognitive resources might be allocated to some later stages of the formation of the stereotype-relevant episodic representations, for example their consolidation. Whereas memory sensitivity (the ability to discriminate between *old* and *new* responses) was not affected by this, bias (the tendency to respond in a particular way) was affected. This effect might indicate that participants could, even under conditions of cognitive load during encoding, discriminate between *old* and *new* items. However, as the episodic representations might have been less consolidated when encoding took place under cognitive load, subsequent memory reconstruction might have been more difficult.

This, in turn, would result in a conservative stereotype-driven bias. The additional activation of stereotypical prototype representations by stereotype-matching cues in the memory test might have challenged the reconstruction of the episodic representations even further, leading participants to favour stereotype-matching responses even more.

The results of Experiment 3 showed that the tendency to respond in a conservative way (old item) was stronger for matching than mismatching questionnaire sentences. This result is in accordance with the source of activation confusion (SAC) model of memory (Diana, Reder, Arndt, & Park, 2006; Reder, Nhouyuaniswong, Schunn, Ayers, Angstadth, & Hiraki, 2000). The model distinguishes between the recollection of a specific encoding event (based on the activation of the episodic representation) and the feeling of familiarity with an item (based on the activation of the semantic representation). Recollection corresponds to the ability to *remember* an item; familiarity corresponds to the feeling of *knowing* an item. For matching items, both recollection and familiarity support to the activation of a stereotype-matching representation. For mismatching items, only recollection supports to the activation of a stereotype-mismatching representation. Therefore, the feeling of knowing in addition to remembering matching items might augment participants' tendency to report an originally matching item as *old.*

## 17. Exploration of the working model within a connectionist approach

An idea for further research would be to express the working model in connectionist terms. The current working model draws its assumptions from associative network, prototype and exemplar models. Its expression in connectionist terms would be a parsimonious way to account for the assumptions from all three models with one mechanism. Although the expression of my working model in connectionist terms does not lead to a novel or different interpretation of my findings, the connectionist approach is the newest approach in the mental representation literature with a strong influence within psychology in general and psycholinguistics in particular (Harley, 2001; Smith, 1998; Smith & DeCoster, 1998). In addition, instantiating my model within a connectionist framework might facilitate future hypothesis generation and testing—for example, about how long a particular representation activation pattern stays activated or how often a reader must process stereotype-mismatching information for the stereotypical representation to be updated. At the moment, I explore the connectionist expression of my working model within one modality. It is therefore restricted to the results of the reading experiments (Experiments 1 to 3). A future challenge would be to describe a connectionist model across modalities.

The reason why the connectionist approach has become popular might be because it is argued to meet limitations of the traditional symbolic approaches (e.g., Clark, 1993; Smith, 1998; Conrey & Smith, 2007). These limitations are shared by my working model, as hybrid of traditional models. Firstly, although contemporary representational theorists agree that representations are better viewed as dynamic than static (e.g., Clark, 1993; Smith, 1998; Conrey & Smith, 2007), my working model has both dynamic and static elements. Viewing representations as dynamic means, for example, that for recollection they are assumed to be reconstructed rather than retrieved as static and unchanged packages. In regards to the approaches contributing to my working model, within exemplar models representations can be viewed as dynamic in a sense that different subsets of exemplars can be activated and therefore different representations constructed online in response to different stimuli in different situations. Within associative networks and prototype models, however, representations are viewed as static in a sense that concepts are activated in an all-or-non fashion (Smith, 1998). Within connectionist models, on the other hand, representations are entirely viewed as transient and dynamic rather than as static knowledge packages.

Secondly, my model, as other traditional models, is a *representation-only model*. That means, that the processes acting upon the representations cannot be inferred from its architecture and extra sets of assumptions have to be made about them (see Smith, 1998). In connectionist models, however, there is no strict distinction between representations and the processes acting upon them, but only a single processing mechanism which means that the architecture of a connectionist network determines how the model processes information.

Thirdly, traditional models have been developed in response to empirical phenomena and are therefore naturally well suited as frameworks to interpret experimental findings (Smith, 1998). However, within these approaches, priority has therefore been given to functionality and no to biological plausibility. Connectionist models, on the other hand, are explicitly oriented on neuronal processes and are therefore often argued to be biologically more plausible than traditional models.

## 17.1 Overview of connectionist models

As a basis for exploring the working model in connectionist terms, I first give a brief (and by no means comprehensive) overview of the architecture and algorithms of different types of connectionist models (for an overview see Harley, 2001; Smith, 1998). Often, within connectionist models, concepts are represented in a distributed way within a network of interconnected units. Each unit within the network has a certain activation level which can change on a fast pace over time (Queller & Smith, 2002). The units are linked by connections with weights attached to them that determine how much activation spreads through them. Different patterns of activation throughout the network represent different concepts. Smith (1998) compares this principle to pixels on a television screen. Whereas the individual pixels have no meaning in themselves, taken together, they can represent many different pictures by taking on diverse patterns of illumination.

In terms of architecture, simple networks include an input layer which is connected to the external input and an output layer which is connected to the external output. The input units' activation pattern corresponds to an external input pattern (e.g., a written word). The output units' activation pattern is related to a specific external output pattern (e.g., conceptual representation of the written word). The activation a unit receives from all the other units connected to it is the sum of the activation levels of those units, multiplied with the weights of the connections. Generally, models can be divided into ones that cannot and ones that can learn. In models that cannot learn (e.g., McClelland & Rumelhart, 1981), the connection weights are predetermined by the developer to obtain from a specific input (e.g., a specific pattern of features) a specific outcome (e.g., a specific word). In models that can learn (e.g., Seidenberg & McClelland, 1989; Chang, Dell, & Bock, 2006), the connection weights are initially set to random by the developer and change over time with training corresponding to a specific learning rule (e.g. back-propagation). One of the advantages of implementing a learning mechanism into a connectionist model is that it can enable the network after training to quite successfully generalise from an initially limited set of input patterns to new inputs (see Harley, 2001; Clark, 1993).  Sadly, the models always need explicit feedback and tuition.

## 17.2 The working model expressed in connectionist terms

In this section, instead of attempting to fully formulate a connectionist model in mathematical terms, I am exploring how the working model could be expressed in connectionist terms. Representations could, contrary to my working model, be distributed. That means that the activation of a particular feature would correspond to the combined activity of several units. The units could be organized in an input layer and an output layer. Input stimuli could be gender-stereotype-relevant words; outputs could be the integrated gender-stereotype-relevant conceptual representations. Any gender or occupation-unrelated words of my sentences could be treated as neutral input. Without restricting myself to a specific learning rule, the formation of new representations, learning and memory could generally be expressed as changes in the connection weights between the network units as a result of the activation patterns evoked by the transformation from input to output.

To represent the knowledge of stereotypes and its changes, a storage technique could be used that Clark (1993) calls superposition: "Two representations are fully superposed if the resources used to represent item 1 are coextensive with those used to represent item 2. Thus, if a network learns to represent item 1 by developing a particular pattern of weights, it will be said to have superposed its representations of items 1 and 2 if it then goes on to encode the information about item 2 by amending the set of original weightings in a way which preserves the functionality (some desired input-output pattern) required to represent item 1 while simultaneously exhibiting the functionality required to represent item 2." (p. 17). Most connectionist models actually make use of partial rather than full superposition, an approach I would also suggest for the connectionist expression of the working model. I would further assume a nonarbitrary construction of the network in a way that semantically similar input would result in a similar activation pattern. This sort of construction would be well suited to the gender-occupation stereotype representation, as it could account for the processes of prototype extraction and generalisation. Clark refers to prototype extraction as “the organisation of knowledge around [such] stereotypical feature sets” (p. 21). The process of prototype extraction starts with encoding a set of exemplars. Smith explains: “An “exemplar representation” may be identified with the set of changes in connection weights produced by the learning mechanism during the processing of a particular stimulus.” If the exemplars have certain feature combinations in common, then the weights of the connections between the common features of all these exemplars would change in a way that the links would become stronger. So, if, for example, many exemplars of female secretaries would have been encoded in the past, then the links for the feature combinations for secretary and female would be strong. Eventually, a prototypical secretary feature set would be generated and any time one feature subset would be activated, the activation of the associated feature subsets would be promoted too (*pattern completion* hereafter). The network could then also generalise to new exemplars, as long as these would share some of the central features of the prototype. As in my working model, prototype extraction could be considered as pre-experimentally set. The new exemplars would be the stereotype-relevant words within my experimental sentences.

The pattern completion property would play a key role in the interpretative functions of the prototype representations. When for example the word *secretary* would be the input, the prototypical pattern of *secretary* would be activated, part of which would be the sub-pattern for *female*. This activation pattern would overlap with the activation pattern for the word *herself* which would facilitate conceptual integration. Would the word *himself* be encountered, however, another activation pattern would be evoked and the conceptual integration would take more time, modelling the mismatch effect. Memory could be expressed as the weights of the connections within the network and learning and representational change could manifest themselves in a change of the connection weights (see Chang et al., 2006; Clark, 1993). This slow process would correspond to the slow statistical learning leading to prototype change in my working model.

## 17.3 Expressing my findings in connectionist terms

The interpretation of my experimental findings within a connectionist approach is speculative in parts and would benefit from a future fully formulated model, allowing implementation and hypotheses testing.

The effect of cognitive load in Experiment 1 could be interpreted in a connectionist way by considering the parallel-constraint-satisfaction mechanism which is an inherent part of connectionist models: the transformation from input to output is constrained by the current input pattern as well as the connection weights, resulting from past learning experiences. The resulting new exemplar representation has a particular activation pattern, defined by a particular combination of unit activation and connection weights. Part of the formation of a new exemplar is assumed to be a conscious and resource-demanding process. Smith specifies: “Resolving inconsistencies and satisfying constraints at the level of consciously accessible knowledge requires effort” (p. 424). I assume that under cognitive load, the transformation of the written input stimuli to the conceptually integrated output concepts can still take place. This would explain the missing effect of cognitive load on the online reading measures. The strength of the exemplar activation pattern might, however, be compromised. If a certain threshold of activation strength could not be met, a quicker fading of the particular activation pattern might be the result. This would equal a decreased consolidation of the representation and could result in it being harder or impossible to be reconstructed later on. In this case, when a participant tries to remember an exemplar, the prototype might be activated via the pattern completion mechanism. This would result in an increased tendency to give a stereotype-matching response. Within a full formulation of the connectionist model, the relationship between resource limitation and activation patterns would have to be specified. It could for example be assumed that there is one set of resources and that if part of it is diverted to remembering a digit, there is less overall activation strength available throughout the rest of the system.

For the interpretation of the results of Experiments 2 and 3 I assume, again, that encountering an occupation label, for example the word *secretary* activates the prototypical pattern for *secretary*. When then the word *herself* is encountered, conceptual integration is facilitated by an overlap of the *female* sub-patterns between *secretary* and *herself*. When next the word *himself* is encountered, no sub-pattern-overlap takes place and conceptual integration would take more time, resulting in a mismatch effect. I assume that upon the encounter of both matching and mismatching pronouns, exemplar representation activation patterns are formed by a change in connection weights. When then another pronoun is encountered, as in Experiment 3, this exemplar activation pattern is assumed to still be active which would make the conceptual integration of the pronoun simple. When, however another occupation label (e.g., *secretary*) is encountered, as in Experiment 2, the prototypical pattern for *secretary*, including the sub-pattern for *female* is reactivated. This most recent activation being stronger than the exemplar pattern activation leads to a repeated mismatch effect on encounter of the next pronoun.

As stated at the outset, I restricted my attempt to interpret my findings within a connectionist framework to one modality. A future full connectionist formulation and implementation might prove fruitful to also model the effects of processing stereotype-relevant information across modalities.

# 18. Broader implications

The results of the series of experiments in this thesis contribute to an understanding of the stability and scope of stereotype-driven processing. The novel methodological approach here was to measure how people integrate stereotype-relevant information into their ongoing processing of social information, in addition to less precise measures (such as questionnaires).

Stereotypes fulfil a number of functions by simplifying and organising social information. They exert an influence on the perception of, and the reasoning about, judgements of and behaviour towards members of stereotyped groups (Hilton & von Hippel, 1996). However, they become problematic when their simplifying character leads to unjustified assumptions about and discriminatory behaviour towards members of the stereotyped group. It has been suggested that a way to change and update stereotypes to reflect reality more accurately is by encountering members of a stereotyped group that mismatch the stereotype (e.g., Rothbart, 1981, see section 7.2). Within my working model, I have assumed, in line with the *bookkeeping model* (Rothbart, 1981), that stereotype update takes place every time mismatching information is encountered. However, this update depends on people noticing, representing and remembering the mismatching information accurately and correctly. The motivation of this thesis was to investigate these processes during reading of stereotype-relevant information.

Going back to the example I used in the introduction about the female labour's deputy leader, I might now be able to answer the questions I posed at the outset. I asked whether readers would even notice this kind of mismatching information when their attention was not undivided (as it rarely is during casual reading). The online findings of Experiment 1 indicate that they still can detect stereotype-mismatches, even when cognitively busy with a concurrent task. The next question I asked was whether readers could remember later on that the deputy leader was a woman. The memory-sensitivity results of Experiments 1 and 3 suggest that they can. Participants did, however, show a bias towards stereotype-matching responses. These results indicate that stereotype-matching representations will not completely outweigh episodic mismatching representations; however, they will still exert an influence on memory, particularly when readers are cognitively busy during encoding, which might impair the consolidation of the mismatching representations to a certain extent.

Another question I asked in the introduction was whether readers would be surprised again when they encountered another piece of stereotype-mismatching information about the same person within the same text. The results of Experiments 2 and 3 suggest that readers do take in the mismatching information and are not surprised again when they encounter more within the same processing episode, unless category-relevant information is reemphasised. These findings show, again, the influence of stereotypes on the processing of social information; however, they also show that this influence is limited, particularly when the category membership of a member of a stereotyped group is not repeated.

My next question was whether reading about the female deputy leader could influence other forms of social information processing, for example of female or male faces.

The results of Experiments 4 to 7 suggest that, indeed, the processing of female and male faces can be influenced by stereotypical information, but only when this information is very salient. Sentences containing stereotype-relevant information did not affect face processing. Whether this finding was due to the linguistic complexity of the materials or to the insensitivity of the task, it again shows the limit of the influence of stereotype activation during reading on further social information processing.

In sum, my research confirms the influence of processing stereotype-relevant information during reading. It also highlights, however, that under certain circumstances, stereotype-mismatching information can outweigh the stereotypical influence. This is particularly interesting because the effects of existing stereotypes on further processing have previously, especially within social psychology, been described as inevitably shaping processing (Bargh, 1999).

# References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73,* 247-264.

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory.* Washington: Winston & Sons.

Ashby, J., & Martin, A. E. (2008). Prosodic phonological representations early in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 224-236.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language, 36,* 94–117.

Bajo, M.-T. (1988). Semantic facilitation with pictures and words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 579-589.

Banaji, M. R. & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science, 7,* 136-141.

Bargh, J. A. (1999). The cognitive monster: The case of automatic stereotype effects. In S. Chaiken and Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361-382). New York: Guildford.

Bargh, J. A. (1994). The four horsemen of automaticity: awareness, intention, efficiency, and control in social cognition. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Handbook of social cognition (2nd ed.): Vol. 1. Basic processes* (pp. 1-40). Hillsdale, NJ: Erlbaum.

Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: the case of information overload. *Journal of Personality and Social Psychology, 49*, 1129–1146.

Baron, I., & Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 386-392.

Baron, J. (1973). Phonemic stage not necessary for reading. *Quarterly Journal of Experimental Psychology, 25,* 241-246.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59,* 617-645.

Belke, E. (2008). Effects of working memory load on lexical access. *Psychonomic Bulletin and Review, 15,* 357-363.

Bierwisch, M., & Schreuder, R. (1992). From concepts to lexical items. *Cognition, 42,* 23-60.

Bower, T. G. R. (1970). Reading by eye. In H. Levin and J. P. Williams (Eds.), *Basic studies on reading* (pp. 134-146). New York: Basic Books.

Brewer, M. B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology, 41*, 656-670.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B.B. Lloyd (Eds.), *Cognition and categorization* (pp. 170-207). New York: Wiley.

Brysbaert, M., Drieghe, D., & Vitu, F. (2005). Word skipping: Implications for theories of eye movement control in reading. In G. Underwood (Ed.), *Cognitive Processes in Eye Guidance* (pp. 53-77). Oxford: Oxford University Press.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology, 14,* 177-208.

Carlston, D. E. (1994). Associated Systems Theory: A systematic approach to the cognitive representation of persons and events. In R. S. Wyer (Ed.), *Advances in Social Cognition: Vol. 7. Associated Systems Theory* (pp. 1-78). Hillsdale, NJ: Erlbaum.

Carlston, D. E., & Smith, E. R. (1996). Principles of mental representation. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles*. New York: Guilford.

Carreiras, M., & Clifton, C., Jr. (2004). On the on-line study of language comprehension. In M. Carreiras and C. Clifton, Jr. (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERP, and beyond* (pp. 1-14). Hove: Psychology Press.

Carreiras, M., Garnham, A., Oakhill, J., & Cain, K. (1996). The use of stereotypical gender information in constructing a mental model: Evidence from English and Spanish. *Quarterly Journal of Experimental Psychology Section A - Human Experimental Psychology, 49,* 639-663.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*, 234-272.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press. Collins.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Clark, A. J. (1993). *Associative Engines: Connectionism, Concepts, and Representational Change*. Cambridge, MA: MIT Press

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359.

Clark, J. M., (1987). Understanding Pictures and Words: Comment on Potter, Kroll, Yachzel, Carpenter, and Sherman (1986). *Journal of Experimental Psychology: General, 116,* 307-309.

Collins, A. M. & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82*, 407-428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240-247.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151-216). London: Academic Press.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100,* 589–608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256.

Conrey, F. R., & Smith, E. R. (2007). Attitude representation: Attitudes as patterns in a distributed, connectionist representational system. *Social Cognition, 25,* 739-758.

Craik, F., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience, 10,* 77-94.

Diana, R. A., & Reder, L. M. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin and Review, 13,* 1-21.

Drieghe, D., Brysbaert, M., & Desmet, T. (2005). Parafoveal-on-foveal effects on eye movements in text reading: Does an extra space make a difference? *Vision research*, *45*, 1693-1706.

Duffy, S. A. & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory and Cognition., 32,* 551-559.

Duffy, S. A., & Rayner, K. (1990). Eye movements and anaphor resolution: Effects of antecedent typicality and distance. *Language and Speech, 33,* 103-119.

Engbert, R., Nuthmann, A., Richter, E. & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112*, 777 – 813.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology, 120,* 339-357.

Federmeier, K. D., & Kutas, M. (2001). Meaning and Modality: Influences of Context, Semantic Memory Organization, and Perceptual Predictability on Picture Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 202-224.

Finkbeiner, M., & Coltheart, M. (2009). Letter recognition: From perception to representation. *Cognitive Neuropsychology, 26*, 1-6.

Fowler, C., Napps, S., & Feldman, L. (1985). Relations among regular and irregular morphologically related words in the lexicon as revealed by repetition priming. *Memory and Cognition, 13,* 241-255.

Frazier, L., & Fodor, J. D. (1978). The Sausage Machine: A New. Two-Stage Parsing Model. *Cognition, 6,* 291-325.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences.
*Cognitive Psychology, 14*, 178-210.

Friederici, A. D., Gunter, T. C., Hahne, A., & Mauth, K. (2004). The relative timing of syntactic and semantic processes in sentence comprehension. *NeuroReport, 15*, 165-169.

Frost, R. (1998). Toward a strong phonological theory of visual word recognition: True issues and false trials. *Psychologycial Bulletin, 123,* 71-99.

Gardiner, J. M., Ramponi, C, & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses. *Memory, 10,* 83-98.

Garnham, A., Oakhill, J., & Reynolds, D. (2002). Are inferences from stereotyped role names to characters' gender made elaboratively? *Memory and Cognition, 30,* 439-446.

Gilboy, E., Sopena, J.-M., Clifton, C. Jr., Frazier, L. (1995). Argument structure and association preferences in Spanish and English complex NPs. *Cognition, 54,* 131-167.

Giraudo, H. & Grainger, J. (2000). Prime word frequency in masked morphological and orthographic priming. *Language and Cognitive Processes, 15,* 421-444.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple readout model. *Psychological Review, 103*, 518–565.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8,* 439–483.

Harley, T. A. (2001). *The psychology of language: From data to theory* (2nd ed.). New York: Psychology Press.

Hashtroudi, S., Mutter, S. A., Cole, E. A., & Green, S. K. (1984). Schema-consistent and schema-inconsistent information: Processing demands. *Personality and Social Psychology Bulletin, 10,* 269–278.

Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology, 37*, 25–38.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment is memory-based or on-line. *Psychological Review, 93,* 258-268.

Hess, D. J., Foss, D. J., & Carroll, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General,124*, 62-92.

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47*, 237-271.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98,* 74-94.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. Psychological review, 93, 411-428.

Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory *(Vol. 18,* pp. 49–94). New York: Academic Press.

Howitt, D., & Cramer, D. (2005). *Introduction to Research Methods in Psychology*. Harlow: Pearson Education.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception and concept learning. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 18, pp. 1-47). New York: Academic Press.

Just, M. A., & Carpenter, P. A. (1992). A Capacity Theory of Comprehension: Individual Differences in Working Memory. *Psychological Review, 99*, 122-149.

Kaiser, E., Runner, J. T, Sussman, R. S., & Tanenhaus, M. K (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition, 112*, 55-80.

Kaiser, E. & Trueswell, J. C. (2004). The Role of discourse context in the processing of a flexible word-order language. *Cognition, 94,* 113-147.

Kawakami, K., & Dovidio, J. F. (2001). Implicit stereotyping: How reliable is it? *Personality and Social Psychology Bulletin*, *27*, 212-225.

Kempley, S. T., & Morton, J. (1982). The effects of priming with regularly and irregularly related words in auditory word recognition. *British Journal of Psychology, 73*, 441-454.

Kennison, S. M., & Trofe, J. L. (2003). Comprehending pronouns: A role for word-specific gender stereotype information. *Journal of Psycholinguistic Research, 32*, 355–378.

Kreiner, H., Sturt, P., & Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language, 58*, 239–261.

Kroll, J. F. (1990). Recognizing words and pictures in sentence contexts: A test of lexical modularity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 747-759.

Le Gal, P. M., & Bruce, V. (2002). Evaluating the independence of sex and expression in judgments of faces. *Perception and Psychophysics, 64,* 230–243.

Lemm, K. M., Dabady, M., & Banaji, M. R. (2005). Gender picture priming: It works with denotative and connotative primes. *Social Cognition, 23*, 218-241.

Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. In *Proceedings for the National Academy of Sciences of the United States of America, 98,* 13464–13471.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22,* 1–75.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distribution of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology, 57*, 165-188.

Liversedge, S. P., Paterson, K. B. Clayes, E. L. (2002). The influence of only syntactic processing “long” relative clause sentences. *The Quarterly Journal of Experimental Psychology, 55A,* 225-240.

Lukatela, G., & Turvey, M. T. (1994). Visual lexical access is initially phonological: 1. Evidence from associative priming by words, homophones, and pseudohomophones. *Journal of Experimental Psychology: General, 123,* 107-128.

Lukatela, G., Gligorijevic, B., Kostic, A., & Turvey, M. T. (1980). Representation of inflected nouns in the internal lexicon. *Memory and Cognition, 8,* 415-23.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 4*, 676-703.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23,* 77-87.

Marshall, J. C., & Newcombe, F. (1973). Patterns of paralexia: A psycholinguistic approach. *Journal of Psycholinguistic Research, 2,* 175-199.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1-71,

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of contexty effects in letter preception: Part 1. An account of the basic findings. *Psychological Review, 88,* 275-407.

McClelland, J. L., St. John. M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes, 4***,** 287-335.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*, 440-466.

McNamara, T. P., & Miller, D. L. (1989). Attributes of theories of meaning. *Psychological Bulletin, 106,* 355-376.

McRae, K., deSa, V. R., & Seidenberg , M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126,* 99–130.

Medin (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 775–799.

Mitchell, D. C. (1994). Sentence parsing. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics (*pp. 375-405). San Diego: Academic Press.

Mullen, B., & Johnson, C. (1995). Cognitive representations in ethnophaulisms and illusory correlation in stereotyping. *Personality and Social Psychology Bulletin, 21*, 420-433.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience, 18*, 1098-1111.

Niswander, E., Pollatsek, A., & Rayner, K. (2000). The processing of derived and inflected suffixed words during reading. *Language and Cognitive Processes, 15*, 389-420.

Nosofsky, R. M. (1986). Attention, similarity, and the identification– categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 924-940.

Oakhill, J., Garnham, A., & Reynolds, D. (2005). Immediate activation of stereotypical gender information. *Memory and Cognition, 33,* 972-983.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 411-421.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language, 34*, 739-773.

Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory and Cognition, 25*, 273-285.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart & Winston.

Paivio, A. (1986). *Mental representations: A dual coding approach.* New York: Oxford University Press.

Perfetti, C. A., Bell, L. C., & Delaney, S. (1988). Automatic (prelexical) phonetic activation in silent word reading: Evidence from backward masking. *Journal of Memory and Language, 27,* 59-70.

Pickering, M. J., Frisson, S., McElree, B., & Traxler, M. J. (2004). Eye-movements and semantic composition. In M. Carreiras, & C.E. Clifton, Jr. (Eds.), *The on-line study of sentence comprehension: Eye-tracking, ERPs, and beyond* (pp. 33-50). Hove: Psychology Press.

Piras, F., & Marangolo, P. (2004). Independent access to phonological and orthographic lexical representations: A replication study. *Neurocase, 10*, 300-307.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 37-42). Hillsdale, NJ: Erlbaum.

Plaut, D. C., & McClelland, J. L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 824–829). Hillsdale, NJ: Erlbaum.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review**,** 103*, 56-115.

Potter, M. C., & Kroll, J. F. (1987). Conceptual representation of pictures and words: Reply to Clark. *Journal of Experimental Psychology: General, 116*, 310-311.

Potter, M. C., Kroll, J. F., Yachzel, B., Carpenter, E., & Sherman, J. (1986). Pictures in sentences: Understanding without words. *Journal of Experimental Psychology: General, 115*, 281-294.

Queller, S., & Smith, E. R. (2002). Subtyping versus bookkeeping in stereotype learning and change: Connectionist simulations and empirical findings. *Journal of Personality and Social Psychology, 82*, 300-313.

Quinn, K. A., & Macrae, C. N. (2005). Categorizing others: The dynamics of person construal. *Journal of Personality and Social Psychology*, *88*, 467–479.

Quinn, K. A., Mason, M. F., & Macrae, C. N. (2009). Familiarity and person construal: Individuating knowledge moderates the automaticity of category activation. *European Journal of Social Psychology, 39*, 852–861.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510-532.

Rayner, K. (1998)**.** Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition, 14*, 191-201.

Rayner, K., & Pollatsek, A. (1989).*The psychology of reading.* New York: Prentice-Hall.

Rayner, K., Juhasz, B. J., & Pollatsek, A. (2005). Eye movements during reading. In M. J. Snowling and C. Hulme (Eds.), *The Science of Reading: A Handbook* (pp. 79-97). Oxford: Blackwell.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 294–320.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105*, 125–157.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*, 445-476.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to Model the Effects of Higher-Level Language Processing on Eye Movements During Reading. *Psychonomic Bulletin and Review, 16*, 1-21.

Reynolds, D. J., Garnham, A., & Oakhill, J. (2006). Evidence of immediate activation of gender information from a social role name. *The Quarterly Journal of Experimental Psychology, 59*, 886–903.

Roelofs, A., Meyer, A. S., & Levelt, W. J. M. (1998). A case for the lemma-lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition, 69*, 219-230.

Rojahn, K., & Pettigrew, T. F. (1992). Memory for schema-relevant information: A meta-analytic resolution. *British Journal of Social Psychology, 31*, 81–109.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Rothbart, M. (1981). Memory processes and social beliefs. In D. L. Hamilton (Ed.), *Cognitive Processes in Stereotyping and Intergroup Behavior* (pp. 145-181). Hillsdale, NJ: Erlbaum.

Rothbart, M., Evans, M., & Fulero, S. (1979). Recall for confirming events: Memory processes and the maintenance of social stereotypes. *Journal of Experimental Social Psychology, 15*, 343-355.

Rossion, B. (2002). Is sex categorization from faces really parallel to face recognition? *Visual Cognition, 9*, 1003–1020.

Schank, R. C. (1972). Conceptual Dependency: A theory of natural language understanding. *Cognitive Psychology, 3,* 532-631.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition. *Psychological Review, 96*, 523-568.

Share, D. L., (2008). On the Anglocentricities of Current Reading Research and Practice: The Perils of Overreliance on an “Outlier” Orthography. *Psychological Bulletin, 134,* 584–615.

Sherman, J. W. (1996). Development and mental representation of stereotypes. *Journal of Personality and Social Psychology, 70*, 1126-41.

Sherman, J. W., & Frost, L. A. (2000). On the encoding of stereotype-relevant information under cognitive load. *Personality and Social Psychology Bulletin, 26,* 26-34.

Sherman, J. W., Conrey, F. R., & Groom, C. J. (2004). Encoding flexibility revisited: Evidence for enhanced encoding of stereotype-inconsistent information under cognitive load. *Social Cognition, 22*, 214-232.

Sherman, J. W., Lee, A. Y., Bessenoff, G. R., & Frost, L. A. (1998). Stereotype efficiency reconsidered: Encoding flexibility under cognitive load. *Journal of Personality and Social Psychology, 75,* 589-606.

Smith, E. R. (1998). Mental representation and memory. In D. Gilbert, S. Fiske & G. Lindzey (Eds.), *Handbook of Social Psychology* (4th ed., Vol. 1, pp. 391-445). New York: McGraw-Hill.

Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology, 74*, 21-35.

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99,* 3-21.

Stangor, C., McMillan, D. (1992). Memory for Expectancy-Congruent and Expectancy-Incongruent Information: A Review of the Social and Social-Developmental Literatures. *Psychological Bulletin, 111,* 42-61.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language, 48,* 542-562.

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical theories of sentence processing. *Cognitive Science, 23*, 491-515.

Taft, M., & Ardasinski, S. (2006). Obligatory decomposition in reading prefixed words. *The Mental Lexicon, 1*, 183-199.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior, 14*, 638-647.

Tanenhaus, M. K, Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension *Science, 268,* 1632-1635.

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier & K. Rayner, (Eds.), *Perspectives in Sentence Processing* (pp. 155-179). Hillsdale, NJ: Erlbaum.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381-403). New York: Academic Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352-373.

Tydgat, I., Grainger, J. (2009). Serial position effects in the identification of letters, digits, and symbols. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 480-498.

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inference. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 28, pp. 211-279). San Diego, CA: Academic Press.

Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience, 11,* 657–671.

Van Gompel, R. P. G., & Pickering, M. J. (in press). Syntactic parsing. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics*. Oxford: Oxford University Press.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review, 97,* 488-522.

Vigliocco, G. & Vinson, D. P. (2007). Semantic Representation. In G. Gaskell (Ed.), *Handbook of Psycholinguistics* (pp. 195-215). Oxford: Oxford University Press.

Vigliocco, G., Vinson, D. P, Lewis, W. & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology, 48,* 422-488.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods, 40,* 183-190.

White, S. J. (2008). Eye movement control during reading: Effects of word frequency and orthographic familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 205-223.

White, S.J., Rayner, K., & Liversedge, S.P. (2005). The influence of parafoveal word length and contextual constraint on fixation durations and word skipping in reading. *Psychonomic Bulletin & Review, 12,* 466-471.

Wiese, H. (1999). *Die Verknüpfung sprachlichen und konzeptuellen Wissens:* Eine Diskussion mentaler Module [The correlation of linguistic and conceptual knowledge: A discussion of mental modules]. In I. Wachsmuth & B. Jung (Eds.)*, KogWis99. Proceedings der 4. Fachtagung der Gesellschaft für Kognitionswissenschaft Bielefeld, 28.September - 1.Oktober 1999*. St. Augustin: Infix-Verlag, 92-97.

Wiese, H. (2004). Semantics as a gateway to language. In H. Härtl & H. Tappe (Eds.), *Mediating between Concepts and Language* (pp. 197-222). Berlin: de Gruyter.

Wilks, Y. (1976). Processing Case. *American Journal of Computational Linguistics, 56, 1-86.*

Yap M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60,* 502-529.

Ziegler**,** J. C., & Jacobs, A. M. (1995). Phonological information provides early sources of constraint in the processing of letter strings. *Journal of Memory and Language, 34,* 567-593.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two Routes or One in Reading Aloud? A Connectionist Dual-Process Model. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1131-1161.

# Appendix

# Appendix 1: Sentence stimuli in Experiment 1

Appendix **Sentences with stereotypically male occupation labels:**

During the journey the pilot injured himself/herself quite badly.

In the evening the mechanic seated himself/herself comfortably in front of the TV.

The article stated that the footballer blamed himself/herself for losing the game.

Often during the day the taxi driver looked at himself/herself in the rear view mirror.

Most of the time the security guard trusted himself/herself to do a good job.

In the afternoon the bricklayer upset himself/herself by damaging the tools.

After work the plumber got himself/herself a big portion of chips.

In the end the carpenter convinced himself/herself that the material was faulty.

Last week the lorry driver almost killed himself/herself driving without lights on.

Quite often the construction worker praised himself/herself for being punctual.

In the evening the butcher washed himself/herself thoroughly and went out.

Every week the locksmith taught himself/herself another little skill.

**Sentences with stereotypically female occupation labels:**

On Monday the babysitter cut herself/himself on a piece of broken glass.

At weekends the nanny was comfortable with herself/himself in the large house.

Many times the housekeeper criticized herself/himself for forgetting birthdays.

Last night the typist introduced herself/himself to the other party guests.

After a while the florist was proud of herself/himself and really liked the job.

At times the childminder asked herself/himself if the children’s diet was right.

On a Sunday the fortune teller treated herself/himself to cakes with cream.

At times the beautician spoke to herself/himself when working alone.

On several occasions the receptionist hurt herself/himself with the sharp scissors.

A month ago the midwife bought herself/himself a new working uniform.

On Saturday the cheerleader dressed herself/himself in a bright costume.

Last week the secretary familiarised herself/himself with the new photocopier.

# Appendix 2: Gender cued-recall questionnaire in Experiment 1

**Questionnaire**

This is a quick test of how well you have remembered the sentences. For each of the following items, please indicate whether the agent was male or female. If you are uncertain please guess. It is important that you answer all the questions.

Thank you!!

| | | | | |
|---|---|---|---|---|
| Was the housekeeper | male | □ | female | □ |
| Was the bricklayer | male | □ | female | □ |
| Was the construction worker | male | □ | female | □ |
| Was the secretary | male | □ | female | □ |
| Was the fortune teller | male | □ | female | □ |
| Was the butcher | male | □ | female | □ |
| Was the mechanic | male | □ | female | □ |
| Was the carpenter | male | □ | female | □ |
| Was the security guard | male | □ | female | □ |
| Was the receptionist | male | □ | female | □ |
| Was the nanny | male | □ | female | □ |
| Was the pilot | male | □ | female | □ |
| Was the locksmith | male | □ | female | □ |
| Was the cheerleader | male | □ | female | □ |
| Was the midwife | male | □ | female | □ |
| Was the beautician | male | □ | female | □ |
| Was the footballer | male | □ | female | □ |
| Was the florist | male | □ | female | □ |
| Was the childminder | male | □ | female | □ |
| Was the plumber | male | □ | female | □ |
| Was the babysitter | male | □ | female | □ |
| Was the taxi driver | male | □ | female | □ |
| Was the typist | male | □ | female | □ |
| Was the lorry driver | male | □ | female | □ |

## Appendix 3: Sentence-memory questionnaire in Experiment 1

Every week the locksmith taught herself/himself

a. another little skill. □

b. to fix another type of lock. □

A month ago the midwife bought himself/herself

a. a new pair of shoes. □

b. a new working uniform. □

In the evening the butcher washed herself/himself

a. thoroughly and went out. □

b. thoroughly and went shopping. □

On a Sunday the fortune teller treated himself/herself

a. to cakes with cream. □

b. to a box of chocolates. □

On Monday the babysitter cut herself/himself

a. on a rusty nail. □

b. on a piece of broken glass. □

After a while the florist was proud of herself/himself

a. and made the best bouquets. □

b. and really liked the job. □

After work the plumber got herself/himself

a. a large pizza. □

b. a big portion of chips. □

At weekends the nanny was comfortable with herself/himself

a. in the large house. □

b. in the little cottage. □

In the evening the mechanic seated himself/herself

a. comfortably on the big sofa. □

b. comfortably in front of the TV. □

On Saturday the cheerleader dressed himself/herself

a. in a bright costume. □

b. in a warm jumper. □

Last week the lorry driver almost killed herself/himself

a. driving without lights on. □

b. overlooking the red light. □

On several occasions the receptionist hurt himself/herself

a. with the sharp scissors. □

b. with the old stapler. □

Last night the typist introduced herself/himself

a. to her new colleagues. □

b. to the other party guests. □

Quite often the construction worker praised herself/himself

a. for being punctual. □

b. for being organised. □

The article stated that the footballer blamed himself/herself

a. for being late for training. □

b. for losing the game. □

At times the childminder asked herself/himself

a. if the children were active enough. □

b. if the children’s diet was right. □

Often during the day the taxi driver looked at himself/herself

a. in the big shop windows. □

b. in the rear view mirror. □

Many times the housekeeper criticized herself/himself

a. for forgetting birthdays. □

b. for forgetting some of the shopping. □

Last week the secretary familiarised himself/herself

a. with the new software. □

b. with the new photocopier. □

In the end the carpenter convinced herself/himself

a. that the material was faulty. □

b. that the drill was broken. □

At times the beautician spoke to himself/herself

a. when working alone. □

b. when driving home from work. □

In the afternoon the bricklayer upset himself/herself

a. by breaking the equipment. □

b. by damaging the tools. □

Most of the time the security guard trusted himself/herself

a. to stay alert. □

b. to do a good job. □

During the journey the pilot injured himself/herself

a. quite badly. □

b. slightly. □

# Appendix 4: Instructions for the load condition of Experiment 1 (no-load condition adapted accordingly)

**Please ensure your mobile phone is turned off, not on silent. The signal can interfere with the equipment.**

This experiment involves reading sentences while wearing an eye tracker and keeping numbers in your memory as well as answering some comprehension questions.

First of all the eye tracker needs to be set up. In order to track your eye movements the eye tracker needs to be fitted by tightening the headband. The experimenter will attempt to make this as comfortable as possible but the headband needs to be tight enough to prevent slipping. Once the eye tracker has been placed on your head please sit as still as possible. Please keep your chin on the chin rest. Try not to nod your head when communicating with the experimenter. The experimenter will set up the cameras to track your eyes correctly. This can be a bit fiddly, but won't involve putting anything in your eyes and isn't painful in any way.

During the experiment, participants sometimes find the eye tracker somewhat heavy or uncomfortable. Please let your experimenter know. The headband can be adjusted.

Once the cameras are set up, you will complete a calibration phase, which involves following a black dot around the screen.

Then the reading and number memorising part of the experiment will begin. Each trial has the following structure:

- At the beginning of each trial a 5 digit number will appear on the screen. Please try to memorise this number. Once you've memorised the number, please press the right button on the back of your push button device.
- A sentence will appear on the screen. Please read the sentence as you would read a book or magazine. Once you have read the sentence, press again the right button on the back of you push button device.
- On some trials you will then get a question about the sentence you have read. The question will have a yes or no answer. If the correct response is **YES,** please press the right button of your push button device. If the correct response is **NO,** please press the left button.
- On other trials another 5 digit number will appear on the screen. Please indicate if that is the same number as the one at the beginning of the trial. If it was the same number**,** please press the right button on the back of your push button device. If it is a different number**,** please press the left button on the back of your screen.

Before the presentation of numbers, sentences and questions, a black dot will appear on the left of the screen. Please look at it until it disappears.

The experiment consists of six blocks. In the short breaks in between take some time to rest your eyes. Blink and perhaps close them for a time. Altogether the experiment will take about 45 minutes.

If you have any questions or concerns, please ask the experimenter now or at any stage during the experiment.

**Thank you for taking part!**

# Appendix 5: ANOVA participant analysis cell means for the agent region in Experiment 1

Means and standard deviations (SD) of the eye-movement measures per condition for the agent area as determined by the participant analyses (N = 29 for half 1, N = 32 for half 2)

| | No-load | | | | Load | | | |
|---|---|---|---|---|---|---|---|---|
| | Match | | Mismatch | | Match | | Mismatch | |
| | Half 1 | Half2 | Half 1 | Half2 | Half 1 | Half2 | Half 1 | Half2 |
| Mean First Fixation Duration | 218 | 217 | 210 | 203 | 204 | 211 | 219 | 217 |
| First Fixation Duration: SD | 42 | 35 | 35 | 27 | 30 | 34 | 31 | 38 |
| Mean First-Pass Duration | 294 | 279 | 284 | 267 | 271 | 256 | 287 | 275 |
| First-Pass Duration: SD | 82 | 70 | 75 | 69 | 81 | 79 | 77 | 74 |
| Mean Selective Regression-Path Duration | 325 | 298 | 324 | 308 | 296 | 289 | 310 | 288 |
| Selective Regression-Path Duration: SD | 102 | 87 | 110 | 80 | 86 | 119 | 86 | 86 |
| Mean Total Reading Time | 409 | 369 | 449 | 384 | 383 | 332 | 467 | 348 |
| Total Reading Time: SD | 176 | 113 | 187 | 120 | 117 | 116 | 185 | 101 |
| Mean Regression Out | 0.17 | 0.12 | 0.17 | 0.18 | 0.19 | 0.19 | 0.14 | 0.14 |
| Regression Out: SD | 0.18 | 0.16 | 0.22 | 0.19 | 0.21 | 0.21 | 0.18 | 0.18 |
| Mean Regression In | 0.19 | 0.20 | 0.25 | 0.20 | 0.21 | 0.19 | 0.32 | 0.14 |
| Regression In: SD | 0.26 | 0.20 | 0.27 | 0.23 | 0.19 | 0.18 | 0.24 | 0.18 |

Note: In all tables in the appendix: First Fixation Duration, First-Pass Duration, Selective Regression-Path Duration, and Total Reading Time are indicated in milliseconds. Regression Out and Regression In are indicated in proportion of valid trials.

# Appendix 6: ANOVA results for the agent region in Experiment 1

| | Factors and interactions | $F_1$ (1,59) | MSE | P | $F_2$ (1,22) | MSE | P |
|---|---|---|---|---|---|---|---|
| First Fixation Duration | Match | .02 | 7.74 | .90 | .03 | 23.79 | .86 |
| | Half | .07 | 50.39 | .79 | .12 | 37.26 | .73 |
| | Load | .02 | 50.57 | .89 | .66 | 228.31 | .43 |
| | Match x Half | 1.30 | 890.68 | .26 | 1.28 | 977.41 | .27 |
| | Match x Load | 15.74 | 7642.38 | .00 | 5.70 | 4658.59 | .03 |
| | Half x Load | .78 | 551.54 | .38 | 2.24 | 859.51 | .15 |
| | Match x Half x Load | .05 | 33.36 | .83 | .10 | 44.30 | .76 |
| First-Pass Duration | Match | .18 | 549.28 | .67 | .49 | 1257.45 | .49 |
| | Half | 2.46 | 13689.69 | .12 | 1.81 | 5758.04 | .19 |
| | Load | .42 | 4852.09 | .52 | 3.26 | 5606.22 | .09 |
| | Match x Half | .00 | 3065.51 | .96 | .01 | 27.02 | .91 |
| | Match x Load | 4.10 | 12459.18 | .05 | 4.79 | 8775.84 | .04 |
| | Half x Load | .02 | 137.19 | .88 | .02 | 69.90 | .88 |
| | Match x Half x Load | .05 | 147.99 | .83 | .05 | 157.90 | .83 |
| Selective Regression-Path Duration | Match | .41 | 1728.90 | .52 | .18 | 622.32 | .68 |
| | Half | 2.95 | 19552.20 | .09 | 2.90 | 8211.99 | .10 |
| | Load | .97 | 19837.24 | .33 | 6.92 | 20148.26 | .02 |
| | Match x Half | .00 | 19.14 | 95 | .13 | 377.15 | .73 |
| | Match x Load | .01 | 48.24 | .92 | .00 | 1.86 | .98 |
| | Half x Load | .12 | 788.50 | .73 | .26 | 1303.35 | .62 |
| | Match x Half x Load | .47 | 2488.98 | .50 | 1.62 | 4594.50 | .22 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| Total Reading Time | Match | 8.30 | 89919.54 | .01 | 11.43 | 65057.78 | .00 |
|  | Half | 19.67 | 285383.22 | .00 | 27.45 | 174877.63 | .00 |
|  | Load | .55 | 25309.57 | .46 | 4.00 | 26603.63 | .06 |
|  | Match x Half | 3.19 | 32503.99 | .08 | 3.51 | 29530.22 | .07 |
|  | Match x Load | .67 | 7303.11 | .42 | 1.00 | 7262.70 | .33 |
|  | Half x Load | 1.11 | 16161.54 | .30 | 1.07 | 10400.76 | .31 |
|  | Match x Half x Load | .68 | 6964.76 | .41 | 1.05 | 12407.28 | .32 |
| Regression Out | Match | .31 | .01 | .58 | .98 | .01 | .33 |
|  | Half | .52 | .00 | .47 | 1.02 | .01 | .32 |
|  | Load | .01 | .00 | .92 | .25 | .01 | .62 |
|  | Match x Half | .48 | .01 | .49 | .45 | .01 | .51 |
|  | Match x Load | 4.35 | .10 | .04 | 3.34 | .05 | .08 |
|  | Half x Load | .15 | .00 | .70 | .26 | .01 | .61 |
|  | Match x Half x Load | .51 | .01 | .48 | 1.10 | .03 | .31 |
| Regression In | Match | 2.24 | .05 | .14 | 2.58 | .06 | .12 |
|  | Half | .6.67 | .20 | .01 | 5.83 | .11 | .02 |
|  | Load | .01 | .00 | .91 | .01 | .00 | .93 |
|  | Match x Half | 5.33 | .17 | .03 | 3.95 | .08 | .06 |
|  | Match x Load | .00 | .00 | .96 | .01 | .00 | .91 |
|  | Half x Load | 3.00 | .09 | .09 | 3.95 | .08 | .06 |
|  | Match x Half x Load | 1.51 | .05 | .22 | .68 | .02 | .42 |

Note: All ANOVA tables contain the results by participants ($F_1$) and items ($F_2$) including mean square error (MSE) and p-values for all factors and interactions

# Appendix 7: ANOVA participant analysis cell means for the pronoun region in Experiment 1

Means and standard deviations (SD) of the eye-movement measures per condition for the pronoun area as determined by the participant analyses (N = 29 for half 1, N = 32 for half 2)

| | No-load | | | | Load | | | |
|---|---|---|---|---|---|---|---|---|
| | Match | | Mismatch | | Match | | Mismatch | |
| | Half 1 | Half2 | Half 1 | Half2 | Half 1 | Half2 | Half 1 | Half2 |
| Mean First Fixation Duration | 205 | 217 | 224 | 225 | 208 | 206 | 225 | 221 |
| First Fixation Duration: SD | 23 | 30 | 43 | 30 | 32 | 40 | 60 | 57 |
| Mean First-Pass Duration | 229 | 243 | 255 | 261 | 230 | 215 | 238 | 239 |
| First-Pass Duration: SD | 48 | 49 | 55 | 49 | 55 | 49 | 64 | 73 |
| Mean Selective Regression-Path Duration | 248 | 254 | 270 | 276 | 242 | 229 | 247 | 258 |
| Selective Regression-Path Duration: SD | 59 | 67 | 67 | 54 | 63 | 54 | 72 | 80 |
| Mean Total Reading Time | 316 | 313 | 418 | 369 | 320 | 303 | 387 | 355 |
| Total Reading Time: SD | 96 | 99 | 196 | 116 | 103 | 83 | 159 | 136 |
| Mean Regression Out | 0.13 | 0.07 | 0.10 | 0.11 | 0.11 | 0.08 | 0.10 | 0.14 |
| Regression Out: SD | 0.18 | 0.13 | 0.12 | 0.14 | 0.18 | 0.13 | 0.17 | 0.21 |
| Mean Regression In | 0.17 | 0.20 | 0.29 | 0.25 | 0.27 | 0.24 | 0.37 | 0.23 |
| Regression In: SD | 0.23 | 0.29 | 0.27 | 0.24 | 0.26 | 0.24 | 0.30 | 0.21 |

# Appendix 8: ANOVA results for the pronoun region in Experiment 1

| | Factors and interactions | $F_1$ (1,59) | MSE | P | $F_2$ (1,22) | MSE | P |
|---|---|---|---|---|---|---|---|
| First Fixation Duration | Match | 12.71 | 12970.36 | .00 | 10.85 | 9699.36 | .00 |
| | Half | .34 | 262.22 | .56 | 1.04 | 974.92 | .32 |
| | Load | .10 | 431.27 | .76 | 1.88 | 1120.39 | .19 |
| | Match x Half | .92 | 650.63 | .34 | .84 | 634.19 | .37 |
| | Match x Load | .07 | 72.41 | .79 | .58 | 463.07 | .45 |
| | Half x Load | 2.11 | 1604.90 | .15 | .02 | 10.87 | .90 |
| | Match x Half x Load | .39 | 273.62 | .54 | .42 | 381.95 | .52 |
| First-Pass Duration | Match | 8.63 | 22116.36 | .01 | 12.50 | 17066.09 | .00 |
| | Half | .08 | 133.16 | .78 | .68 | 1718.32 | .42 |
| | Load | 2.48 | 16733.34 | .12 | 10.47 | 14454.03 | .00 |
| | Match x Half | .11 | 176.61 | .75 | .08 | 135.42 | .78 |
| | Match x Load | .18 | 471.86 | .67 | .21 | 589.30 | .65 |
| | Half x Load | 2.57 | 4190.93 | .11 | .02 | 23.82 | .88 |
| | Match x Half x Load | 1.33 | 2207.59 | .25 | 1.48 | 3647.84 | .24 |
| Selective Regression-Path Duration | Match | 6.78 | 21995.24 | .01 | 9.90 | 17692.34 | .01 |
| | Half | .21 | 398.27 | .65 | .67 | 2819.35 | .42 |
| | Load | 2.02 | 19668.48 | .16 | 13.57 | 19583.03 | .00 |
| | Match x Half | .96 | 2139.50 | .33 | .30 | 526.44 | .59 |
| | Match x Load | .13 | 414.04 | .72 | .16 | 455.71 | .69 |
| | Half x Load | .35 | 666.01 | .56 | .22 | 264.79 | .64 |
| | Match x Half x Load | .96 | 2140.24 | .33 | 1.11 | 3938.74 | .30 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| Total Reading Time | Match | 23.98 | 292908.21 | .00 | 34.98 | 258002.04 | .00 |
|  | Half | 5.08 | 38818.39 | .03 | 2.02 | 21847.09 | .17 |
|  | Load | .24 | 9240.33 | .63 | 2.34 | 17185.92 | .14 |
|  | Match x Half | 1.92 | 13590.06 | .17 | 3.33 | 19400.85 | .08 |
|  | Match x Load | .48 | 5898.48 | .49 | 1.27 | 10143.44 | .27 |
|  | Half x Load | .00 | 17.52 | .96 | .33 | 3342.40 | .57 |
|  | Match x Half x Load | .50 | 3497.61 | .48 | .90 | 6457.07 | .35 |
| Regression Out | Match | .24 | .01 | .63 | .72 | .01 | .40 |
|  | Half | .44 | .01 | .51 | .10 | .00 | .76 |
|  | Load | .07 | .00 | .79 | .01 | .00 | .93 |
|  | Match x Half | 5.81 | .07 | .02 | 3.45 | .03 | .08 |
|  | Match x Load | .38 | .01 | .54 | .00 | .00 | .97 |
|  | Half x Load | .60 | .01 | .44 | .12 | .00 | .73 |
|  | Match x Half x Load | .00 | .00 | .99 | .03 | .00 | .86 |
| Regression In | Match | 6.62 | .25 | .01 | 8.48 | .28 | .01 |
|  | Half | 2.68 | .11 | .11 | 2.21 | .05 | .15 |
|  | Load | .97 | .14 | .33 | 1.98 | .06 | .17 |
|  | Match x Half | 3.86 | .13 | .05 | 3.84 | .10 | .06 |
|  | Match x Load | .68 | .03 | .41 | .75 | .02 | .40 |
|  | Half x Load | 2.46 | .10 | .12 | 2.66 | .08 | .12 |
|  | Match x Half x Load | .25 | .01 | .62 | .00 | .00 | .99 |

# Appendix 9: ANOVA participant analysis cell means for the pronoun spill-over region in Experiment 1

Means and standard deviations (SD) of the eye-movement measures per condition for the pronoun spill-over area as determined by the participant analyses (N = 29 for half 1, N = 32 for half 2)

| | No-load | | | | Load | | | |
|---|---|---|---|---|---|---|---|---|
| | Match | | Mismatch | | Match | | Mismatch | |
| | Half 1 | Half2 | Half 1 | Half2 | Half 1 | Half2 | Half 1 | Half2 |
| Mean First Fixation Duration | 232 | 231 | 240 | 227 | 225 | 241 | 222 | 230 |
| First Fixation Duration: SD | 38 | 48 | 43 | 56 | 47 | 50 | 52 | 44 |
| Mean First-Pass Duration | 278 | 289 | 299 | 277 | 264 | 276 | 290 | 281 |
| First-Pass Duration: SD | 58 | 97 | 84 | 72 | 61 | 69 | 126 | 66 |
| Mean Selective Regression-Path Duration | 303 | 334 | 330 | 317 | 286 | 311 | 353 | 317 |
| Selective Regression-Path Duration: SD | 69 | 122 | 97 | 91 | 70 | 92 | 133 | 97 |
| Mean Total Reading Time | 414 | 396 | 441 | 383 | 384 | 372 | 432 | 381 |
| Total Reading Time: SD | 150 | 125 | 187 | 140 | 137 | 126 | 167 | 127 |
| Mean Regression Out | 0.10 | 0.19 | 0.21 | 0.21 | 0.21 | 0.21 | 0.37 | 0.29 |
| Regression Out: SD | 0.18 | 0.26 | 0.21 | 0.25 | 0.25 | 0.24 | 0.32 | 0.26 |
| Mean Regression In | 0.26 | 0.20 | 0.23 | 0.15 | 0.24 | 0.16 | 0.21 | 0.17 |
| Regression In: SD | 0.22 | 0.21 | 0.28 | 0.17 | 0.30 | 0.22 | 0.27 | 0.22 |

# Appendix 10: ANOVA results for the pronoun spill-over region in Experiment 1

| | Factors and interactions | $F_1$ (1,59) | MSE | P | $F_2$ (1,21) | MSE | P |
|---|---|---|---|---|---|---|---|
| First Fixation Duration | Match | .33 | 323.78 | .57 | .02 | 11.10 | .90 |
| | Half | .22 | 404.88 | .64 | .00 | 1.55 | .97 |
| | Load | .12 | 594.35 | .73 | 1.18 | 1129.14 | .29 |
| | Match x Half | 1.24 | 1421.69 | .27 | .55 | 593.77 | .47 |
| | Match x Load | 1.37 | 1343.01 | .25 | 5.68 | 5808.77 | .03 |
| | Half x Load | 3.15 | 5902.30 | .08 | 3.73 | 3753.75 | .07 |
| | Match x Half x Load | .04 | 47.99 | .84 | .28 | 383.23 | .61 |
| First-Pass Duration | Match | 1.61 | 6027.48 | .21 | 5.25 | 7134.15 | .03 |
| | Half | .04 | 240.83 | .84 | .41 | 1122.21 | .53 |
| | Load | .33 | 3773.01 | .57 | 2.03 | 5405.71 | .17 |
| | Match x Half | .08 | 462.04 | .18 | .59 | 1518.69 | .45 |
| | Match x Load | .53 | 1989.20 | .47 | .87 | 3229.01 | .36 |
| | Half x Load | .11 | 613.13 | .75 | 3.44 | 7126.00 | .08 |
| | Match x Half x Load | .08 | 462.04 | .78 | .34 | 1311.71 | .57 |
| Selective Regression-Path Duration | Match | 3.48 | 25886.03 | .07 | 5.29 | 21210.07 | .03 |
| | Half | .02 | 156.22 | .89 | .25 | 1298.80 | .62 |
| | Load | .07 | 1002.64 | .79 | 1.39 | 4696.33 | .25 |
| | Match x Half | 4.27 | 42507.35 | .04 | 2.26 | 8423.44 | .15 |
| | Match x Load | 1.94 | 14431.31 | .17 | .37 | 1343.18 | .55 |
| | Half x Load | .40 | 3096.99 | .53 | .34 | 1578.78 | 57 |
| | Match x Half x Load | .12 | 1153.35 | .74 | 1.30 | 5795.22 | .27 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Total Reading Time | Match | 1.75 | 19533.98 | .19 | .62 | 7002.59 | .44 |
| | Half | 4.19 | 74346.75 | .05 | .4.31 | 3373.48 | .05 |
| | Load | .34 | 14935.38 | .56 | .82 | 6926.35 | .38 |
| | Match x Half | 1.84 | 23559.68 | .18 | .59 | 3303.27 | .45 |
| | Match x Load | .63 | 7010.73 | .43 | 1.05 | 5164.61 | .32 |
| | Half x Load | .04 | 679.48 | .85 | .17 | 1661.36 | .68 |
| | Match x Half x Load | .00 | 6.56 | .98 | .33 | 3467.46 | .57 |
| Regression Out | Match | 10.52 | .50 | .00 | 10.53 | .33 | .00 |
| | Half | .01 | .00 | .92 | 1.75 | .03 | .20 |
| | Load | 5.96 | .56 | .02 | 10.94 | .38 | .00 |
| | Match x Half | 2.01 | .11 | .16 | .26 | .01 | .61 |
| | Match x Load | 1.05 | .05 | .31 | .05 | .00 | .83 |
| | Half x Load | 1.85 | .10 | .18 | .90 | .02 | .36 |
| | Match x Half x Load | .02 | .00 | .90 | .89 | .03 | .36 |
| Regression In | Match | .70 | .04 | .41 | 1.36 | .06 | .26 |
| | Half | 6.78 | .25 | .01 | 4.66 | .17 | .04 |
| | Load | .20 | .02 | .66 | .15 | .00 | .70 |
| | Match x Half | .14 | .01 | .71 | .21 | .00 | .65 |
| | Match x Load | .20 | .01 | .66 | .92 | .03 | .35 |
| | Half x Load | .05 | .00 | .82 | .59 | .01 | .45 |
| | Match x Half x Load | .33 | .02 | .57 | .03 | .00 | .87 |

# Appendix 11: ANOVA participant analysis cell means for the entire sentence region in Experiment 1

Mean Sentence Reading Time (ms) and Sentence Fixation Counts (number of fixations) with standard deviations (SD) as determined by the participant analyses (N = 29 for half 1, N = 32 for half 2)

| | No-load | | | | Load | | | |
|---|---|---|---|---|---|---|---|---|
| | Match | | Mismatch | | Match | | Mismatch | |
| | Half 1 | Half2 | Half 1 | Half2 | Half 1 | Half2 | Half 1 | Half2 |
| Mean Sentence Reading Time | 3159 | 2863 | 3431 | 2915 | 2773 | 2392 | 2974 | 2560 |
| Sentence Reading Time: SD | 877 | 778 | 924 | 755 | 724 | 670 | 749 | 732 |
| Mean Sentence Fixation Count | 14.08 | 13.13 | 15.16 | 13.47 | 12.34 | 10.70 | 13.17 | 11.41 |
| Sentence Fixation Count: SD | 3.79 | 3.36 | 4.23 | 3.23 | 2.67 | 2.46 | 2.96 | 2.53 |

# Appendix 12: ANOVA results for the sentence region in Experiment 1

| | Factors and interactions | $F_1$ (1,59) | MSE | P | $F_2$ (1,22) | MSE | P |
|---|---|---|---|---|---|---|---|
| Sentence Reading Time | Match | 22.43 | 1824661.6 | .00 | 25.64 | 1539295.6 | .00 |
| | Half | 63.04 | 9817515.2 | .00 | 60.03 | 7131981.9 | .00 |
| | Load | 5.02 | 10586382 | .03 | 73.08 | 7147023.4 | .00 |
| | Match x Half | 3.56 | 240302.96 | .06 | 1.74 | 148831.64 | .20 |
| | Match x Load | .09 | 7245.32 | .77 | .01 | 1902.57 | .90 |
| | Half x Load | .01 | 1084.05 | .93 | .01 | 2431.02 | .91 |
| | Match x Half x Load | 1.96 | 132566.92 | .17 | .02 | 6115.63 | .88 |
| Sentence Fixation Count | Match | 19.12 | 33.27 | .00 | 16.66 | 29.14 | .00 |
| | Half | 52.83 | 138.63 | .00 | 37.98 | 101.02 | .00 |
| | Load | 7.36 | 256.07 | .01 | 81.72 | 175.93 | .00 |
| | Match x Half | 2.07 | 2.76 | .16 | .81 | 1.17 | .38 |
| | Match x Load | .04 | .06 | .85 | .03 | .04 | .86 |
| | Half x Load | .85 | 2.23 | .36 | 1.60 | 3.11 | .22 |
| | Match x Half x Load | 1.09 | 1.46 | .30 | .02 | .06 | .89 |

## Appendix 13: Pretest questionnaire for the item *goalkeeper* in Experiment 2

### Questionnaire - Occupation Rating

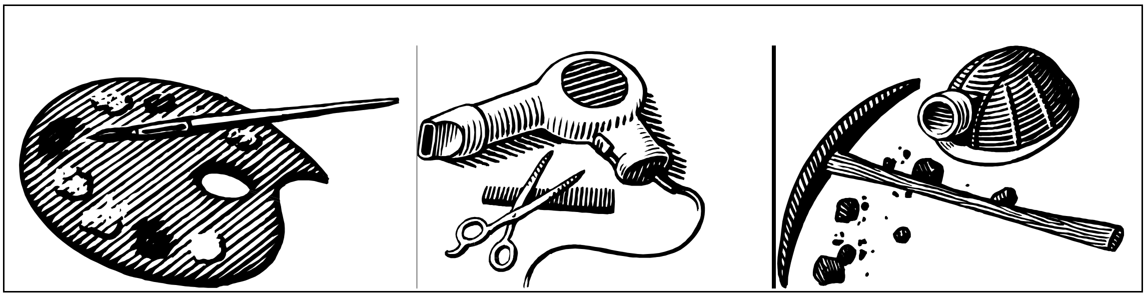
**Dear participant,**

We are interested in how typically male or female you regard the occupations listed below. Please read the examples of occupations, and circle the response that YOU feel is appropriate. There are no right or wrong answers. It is your perceptions that we are interested in. Please provide an answer to **all** the questions listed by ticking the appropriate number on the following scale:

**1 -------------- 2 -------------- 3 -------------- 4 -------------- 5 -------------- 6 -------------- 7**

**Not at all** **Very much so**

If you have any questions please ask them now.
Thank you.



**Chef**

**Do you regard this occupation as typically male?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all** **Very much so**

**Do you regard this occupation as typically female?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all** **Very much so**

---

**Typist**

**Do you regard this occupation as typically male?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all** **Very much so**

**Do you regard this occupation as typically female?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all** **Very much so**

---

**Lawyer**

**Do you regard this occupation as typically male?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all** **Very much so**

**Do you regard this occupation as typically female?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all** **Very much so**

---

**Goalkeeper**

**Do you regard this occupation as typically male?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all Very much so**

**Do you regard this occupation as typically female?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all Very much so**

**Model**

**Do you regard this occupation as typically male?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all Very much so**

**Do you regard this occupation as typically female?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all Very much so**

**Decorator**

**Do you regard this occupation as typically male?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all Very much so**

**Do you regard this occupation as typically female?**

**1 --------------- 2 --------------- 3 --------------- 4 --------------- 5 --------------- 6 --------------- 7**
**Not at all Very much so**

## Appendix 14: Experimental sentence stimuli in Experiment 2: token condition (1) and type condition (2)

(1) Last week, the drunken lorry driver almost killed himself/herself driving through a red light and really scared an old man on the footpath. In addition, the lorry driver completely embarrassed himself/herself by not knowing the route to Cardiff.
(2) Last week, the drunken lorry driver almost killed himself/herself driving through a red light and severely injured an old man on the footpath. Unfortunately, the replacement lorry driver completely embarrassed himself/herself by not knowing the route to Cardiff.

(1) For the placement in Africa, the English midwife bought herself/himself a new uniform, along with some gifts for the local colleague, who had planned the visit, including some walks. The midwife had already ordered some sturdy shoes for himself, which would be useful for the hikes.
(2) For the placement in Africa, the English midwife bought herself/himself a new uniform, along with some gifts for the local colleague, who had planned the visit, including some walks. The African midwife had already ordered some sturdy shoes for herself/himself, which would be useful for the hikes.

(1) The article stated that the goalkeeper blamed himself/herself for losing the game and decided to take a short break from the team. In the new season, the goalkeeper promised to devote himself/herself completely to training.
(2) The article stated that the goalkeeper blamed himself/herself for losing the game and was told by the coach to leave the team. The new goalkeeper promised to devote himself/herself completely to training.

(1) After the final session, the famous fortune teller treated herself/himself to cakes with cream brought in by a close friend. In spite of some hate mail, the fortune teller thought a great deal of herself/himself for providing everyone with sound advice.
(2) After the final session, the famous fortune teller treated himself/herself to cakes with cream brought in by a less successful colleague. In spite of some hate mail, this fortune teller thought a great deal of himself/herself for providing everyone with sound advice.

(1) In the past, the young construction worker had often praised himself/herself for being punctual and for covering for an older workmate who frequently ran late. However, lately the young construction worker had allowed himself/herself the luxury of being a little late as well.
(2) In the past, the young construction worker had often praised himself/herself for being punctual and for covering for an older workmate who frequently ran late. However, lately the older construction worker had allowed himself/herself the luxury of being late a bit too frequently.

(1) Every week, the locksmith taught himself/herself a new skill using a handbook written by an American expert. Through this routine the locksmith quickly established himself/herself as particularly competent.
(2) Every week, the locksmith taught himself/herself a new skill using a handbook written by an American expert. Through this handbook, the American locksmith quickly established himself/herself as particularly competent.

(1) Last week, the babysitter cut herself/himself on a piece of broken glass and almost fainted before the children's eyes.  In spite of the injury, the babysitter forced herself/himself to read to the children until the parents returned.
(2) Last week, the babysitter cut herself/himself on a piece of broken glass and fainted before the children's eyes. In spite of heavy migraines, the neighbour's babysitter forced herself/himself to read to the children until the parents returned.

(1) In the evening, the young mechanic seated herself/himself comfortably in front of the TV and watched the all-night song contest with an old friend. At bedtime, the young mechanic found it difficult to drag herself/himself away from the program.

(2) In the evening, the young mechanic seated himself/herself comfortably in front of the TV and watched the all-night song contest with an old colleague. At bedtime, the older mechanic found it difficult to drag himself/herself away from the program.

(1) On Saturday, the cheerleader dressed himself/herself in a smart outfit and had lunch with an elderly neighbour who had recently returned from hospital. Being new to the area, the cheerleader still struggled to establish himself/herself in the quiet village.
(2) On Saturday, the cheerleader dressed herself/himself in a smart outfit and had lunch with a friend who had recently joined the team. Having only just started, the new cheerleader still struggled to establish herself/himself in the new team.

(1) The overworked security guard trusted himself/herself to do a good job but had overlooked several suspicious parcels and was criticised by the supervisor. Therefore the security guard had to acquaint himself/herself with the complicated regulations again.
(2) The overworked security guard trusted himself/herself to do a good job but had overlooked several suspicious parcels and was dismissed by the supervisor. The new security guard had to acquaint himself/herself with the complicated regulations.

(1) Several times, the younger of the two receptionists had hurt herself/himself with the scissors but had never needed any help. However, this time the cut was deep and the receptionist could barely keep herself/himself from fainting.
(2) Several times, the younger of the two receptionists had hurt herself/himself with the scissors, but had never needed help. However, this time, the cut was deep and the other receptionist could barely keep herself/himself from fainting.

(1) A dangerous habit of the taxi driver was to look at himself/herself in the rear view mirror, which was reported by an Italian colleague. Nevertheless, the taxi driver did not consider himself/herself to be particularly irresponsible.
(2) A dangerous habit of the taxi driver was to look at himself/herself in the rear view mirror, which was reported by an Italian colleague. Naturally, the Italian taxi driver did not consider himself/herself to be quite as irresponsible.

(1) After a while, the florist was proud of herself/himself and liked the job in spite of the grumpy colleague working in the greenhouse. Nevertheless, the enthusiastic florist thought of going into business for herself/himself as soon as possible.
(2) After a while, the florist was proud of herself/himself and liked the job in spite of the grumpy colleague working in the greenhouse. Luckily, the irritable florist thought of going into business for herself/himself as soon as possible.

(1) In the evening, the butcher washed himself/herself thoroughly and visited the village fair with a neighbour. However, the butcher did not enjoy himself/herself there and went home early.
(2) In the evening, the butcher washed himself/herself thoroughly and visited the village fair with a new colleague. However, the new butcher did not enjoy himself/herself there and went home early.

(1) At times, the trainee childminder asked herself/himself whether the children’s diet was right and finally decided to consult an experienced nutritionist. Previously the childminder had only set herself/himself the target of providing a bit of fresh fruit every day.
(2) At times, the trainee childminder asked herself/himself whether the children’s diet was right and finally decided to consult the experienced trainer. This qualified childminder had only set herself/himself the target of providing a bit of fresh fruit every day.

(1) During the journey, the experienced pilot injured himself/herself quite badly and was told by the doctor to take a long holiday. After returning to the job, the pilot had to familiarise himself/herself with the cockpit again.
(2) During the journey, the experienced pilot injured himself/herself quite badly and was told by the doctor to take a long holiday. The next day, a younger pilot/herself had to familiarise himself with the cockpit.

(1) Throughout the years, the housekeeper had often criticized herself/himself for forgetting birthdays and finally asked a friend for advice. From then on, the housekeeper used a calendar to remind herself/himself of important dates.

(2) Throughout the years, the housekeeper had often criticized herself/himself for forgetting birthdays and finally asked a friend for advice. Not surprisingly, the other housekeeper used a calendar to remind herself/himself of important dates.

(1) The elderly secretary thoroughly familiarised herself/himself with the new computer a few months before retiring. To everyone’s surprise, the secretary really enjoyed herself/himself while exploring the potential of the computer.
(2) The elderly secretary reluctantly familiarised herself/himself with the new computer a few months before retiring. In contrast, the new secretary really enjoyed herself/himself while exploring the potential of the computer.

(1) In the end, the carpenter convinced himself/herself that the material was indeed faulty, as suspected by a Swiss colleague. In fact, the carpenter had never regarded himself/herself as an expert.
(2) In the end, the carpenter convinced himself/herself that the material was indeed faulty, as suspected by a Swiss colleague. Surprisingly, the Swiss carpenter had never regarded himself/herself as an expert.

(1) In the afternoon, the bricklayer upset himself/herself by damaging the tools and was asked by the foreman to consider further training. However, the bricklayer decided to restrict himself/herself to less demanding jobs.
(2) In the afternoon, the bricklayer upset himself/herself by damaging the tools and was asked by the foreman to leave and get further training. The new bricklayer decided to restrict himself/herself to less demanding jobs.

(1) Last night, the typist introduced herself/himself to the guests at the company party as the new member of the administrative team. Only a few hours earlier, the typist had excused herself/himself from another party.
(2) Last night, the typist introduced herself/himself to the guests at the company party as the only member of the administrative team. A few hours earlier, the other typist had excused herself/himself from attending.

(1) After work, the plumber got himself/herself a big portion of chips even though the doctor had strongly recommended a low-fat diet. The hungry plumber was unable to control himself/herself when it came to chips.
(2) After work, the plumber got himself/herself a big portion of chips with a German colleague who claimed to be on a low-fat diet. The German plumber was unable to control himself/herself when it came to chips.

(1) With the other staff around, the nanny was comfortable with herself/himself in the large house, but did not like being there alone on weekends. Finally the family posted an advert for a weekend replacement and the nanny counted herself/himself lucky to have the weekends off.
(2) With the other staff around, the nanny was comfortable with herself/himself in the large house, but did not like being there alone on weekends. Finally the family posted an advert for a weekend replacement and the new nanny counted herself/himself lucky to cover the shifts and earn some extra money.

(1) One morning, the beautician spoke aloud to herself/himself about serious family problems without realising that the receptionist was listening from the next room. The unhappy beautician was deeply ashamed of herself/himself on learning that the receptionist was gossiping about these problems.
(2) One morning, the beautician spoke aloud to herself/himself about serious family problems without realising that a young colleague was listening from the next room. This recently hired beautician was deeply ashamed of herself/himself when caught gossiping about these problems.

# Appendix 15: Overview of the design of Experiment 2

| | | Version 1 | Version 2 | Version 3 | Version 4 |
|---|---|---|---|---|---|
| | **Agent** | **Order 1/Order 2** | **Order 1/Order 2** | **Order 1/Order 2** | **Order 1/Order 2** |
| 1 | Babysitter | match/token | mismatch/token | match/type | mismatch/type |
| 2 | Nanny | | | | |
| 3 | Housekeeper | | | | |
| 4 | Typist | match/type | mismatch/type | mismatch/token | match/token |
| 5 | Florist | | | | |
| 6 | Childminder | | | | |
| 7 | Fortune teller | mismatch/token | match/token | mismatch/type | match/type |
| 8 | Receptionist | | | | |
| 9 | Midwife | | | | |
| 10 | Beautician | mismatch/type | match/type | match/token | mismatch/token |
| 11 | Cheerleader | | | | |
| 12 | Secretary | | | | |
| 13 | Pilot | match/token | mismatch/token | match/type | mismatch/type |
| 14 | Mechanic | | | | |
| 15 | Goalkeeper | | | | |
| 16 | Security guard | match/type | mismatch/type | mismatch/token | match/token |
| 17 | Bricklayer | | | | |
| 18 | Plumber | | | | |
| 19 | Taxi Driver | mismatch/token | match/token | mismatch/type | match/type |
| 20 | Carpenter | | | | |
| 21 | Lorry driver | | | | |
| 22 | Butcher | mismatch/type | match/type | match/token | mismatch/token |
| 23 | Construction worker | | | | |
| 24 | Locksmith | | | | |

# Appendix 16: Instructions for Experiment 2 and 3

Hello!

In this experiment you will be presented with sentences. Please read them quietly and move on by pressing the right button on the back of the push button device. After some of the sentences you will be asked questions. For responding with 'yes' please press the right button on the back of the push button device, for 'no' please press the left button.

If you have any questions, please ask your experimenter now. Otherwise press any button to start the experiment.

Thanks!

## Appendix 17: ANOVA participant analysis cell means for the agent region in Experiment 2

Cell means (RT) and standard deviations (SD) of the eye-movement measures for the agent interest area (in msec and number of regressions) as determined by the participant analyses in the ANOVA (factors: token-type (token/type), match (match/mismatch), sentence number (sen 1/sen 2); N = 40 for sen 1, N = 40 for sen 2)

| | Token | | | | Type | | | |
|---|---|---|---|---|---|---|---|---|
| | Match | | Mismatch | | Match | | Mismatch | |
| | Sen 1 | Sen 2 | Sen 1 | Sen 2 | Sen 1 | Sen 2 | Sen 1 | Sen 2 |
| Mean First Fixation Duration | 245 | 239 | 248 | 237 | 249 | 237 | 248 | 235 |
| First Fixation Duration: SD | 45 | 42 | 40 | 48 | 34 | 39 | 43 | 37 |
| Mean First-Pass Duration: RT | 335 | 306 | 342 | 299 | 350 | 290 | 368 | 287 |
| First-Pass Duration: SD | 86 | 71 | 81 | 86 | 78 | 63 | 124 | 74 |
| Mean Selective Regression-Path Duration | 357 | 325 | 365 | 310 | 372 | 304 | 389 | 302 |
| Selective Regression-Path Duration: SD | 77 | 71 | 88 | 87 | 87 | 62 | 124 | 82 |
| Mean Total Reading Time | 435 | 363 | 460 | 369 | 449 | 361 | 488 | 356 |
| Total Reading Time: SD | 140 | 90 | 163 | 103 | 111 | 95 | 165 | 97 |
| Mean Regression Out | 0.12 | 0.11 | 0.11 | 0.10 | 0.09 | 0.13 | 0.11 | 0.14 |
| Regression Out: SD | 0.15 | 0.16 | 0.16 | 0.17 | 0.14 | 0.14 | 0.14 | 0.19 |
| Mean Regression In | 0.14 | 0.07 | 0.18 | 0.15 | 0.10 | 0.12 | 0.17 | 0.07 |
| Regression In: SD | 0.21 | 0.10 | 0.25 | 0.14 | 0.15 | 0.17 | 0.18 | 0.12 |

## Appendix 18: ANOVA results for the agent region in Experiment 2

| | Factors and interactions | $F_1$ (1,39) | MSE | P | $F_2$ (1,23) | MSE | P |
|---|---|---|---|---|---|---|---|
| First Fixation Duration | Match | .05 | 39.73 | .82 | .01 | 7.79 | .93 |
| | Sentence Number | 7.66 | 8308.93 | .01 | 5.65 | 6580.43 | .03 |
| | Token-Type | .00 | 1.02 | .97 | .00 | .33 | .99 |
| | Match x Sentence Number | .24 | 207.92 | .63 | .03 | 16.13 | .86 |
| | Match x Token-Type | .02 | 27.80 | .90 | .00 | 2.11 | .97 |
| | Sentence Number x Token-Type | .16 | 222.18 | .70 | .01 | 5.82 | .94 |
| | Match x Sentence Number x Token-Type | .05 | 65.72 | .82 | .00 | .10 | .99 |
| First-Pass Duration | Match | .14 | 1106.63 | .71 | .19 | 847.90 | .67 |
| | Sentence Number | 38.80 | 224865.24 | .00 | 41.45 | 133645.30 | .00 |
| | Token-Type | .14 | 700.69 | .72 | .00 | 5.14 | .97 |
| | Match x Sentence Number | 1.52 | 6547.13 | .23 | .80 | 1662.75 | .38 |
| | Match x Token-Type | .19 | 998.07 | .67 | .01 | 19.87 | .93 |
| | Sentence Number x Token-Type | 5.92 | 24569.35 | .02 | 3.30 | 10869.11 | .08 |
| | Match x Sentence Number x Token-Type | .07 | 249.89 | .80 | .37 | 806.96 | .55 |
| Selective Regression-Path Duration | Match | .03 | 258.55 | .86 | .07 | 267.48 | .80 |
| | Sentence Number | 54.77 | 291163.44 | .00 | 44.64 | 180228.16 | .00 |
| | Token-Type | .11 | 594.00 | .75 | .00 | 13.32 | .95 |
| | Match x Sentence Number | 2.24 | 9100.44 | .14 | .65 | 1829.65 | .43 |
| | Match x Token-Type | .39 | 2240.32 | .54 | .12 | 349.33 | .73 |
| | Sentence Number x Token-Type | 4.70 | 22986.91 | .04 | 4.19 | 9533.58 | .05 |
| | Match x Sentence Number x Token-Type | .01 | 66.50 | .91 | .06 | 198.33 | .80 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Total Reading Time | Match | 1.74 | 20534.44 | .20 | 1.77 | 9132.57 | .20 |
| | Sentence Number | 58.26 | 733035.25 | .00 | 52.40 | 454453.27 | .00 |
| | Token-Type | .20 | 3460.95 | .66 | .12 | 1111.78 | .73 |
| | Match x Sentence Number | 1.59 | 19998.39 | .22 | 1.11 | 6196.43 | .30 |
| | Match x Token-Type | .00 | 49.25 | .95 | .02 | 133.53 | .88 |
| | Sentence Number x Token-Type | 2.20 | 17174.78 | .15 | 3.62 | 14244.04 | .07 |
| | Match x Sentence Number x Token-Type | .32 | 3441.24 | .58 | 1.03 | 5202.92 | .32 |
| Regression Out | Match | .09 | .00 | .76 | .05 | .00 | .83 |
| | Sentence Number | .59 | .01 | .45 | .01 | .00 | .91 |
| | Token-Type | .04 | 00 | .83 | .37 | .00 | .55 |
| | Match x Sentence Number | .01 | .00 | .92 | .00 | .00 | .99 |
| | Match x Token-Type | .29 | .01 | .60 | .26 | .01 | .61 |
| | Sentence Number x Token-Type | 1.80 | .03 | .19 | 1.58 | .02 | .22 |
| | Match x Sentence Number x Token-Type | .04 | .00 | .85 | .05 | .00 | .83 |
| Regression In | Match | 4.14 | .09 | .05 | 3.26 | .04 | .08 |
| | Sentence Number | 5.18 | .18 | .03 | 6.74 | .10 | .02 |
| | Token-Type | 1.37 | .03 | .25 | .77 | .01 | .39 |
| | Match x Sentence Number | 1.31 | .04 | .26 | .88 | .02 | .36 |
| | Match x Token-Type | 3.53 | .05 | .07 | 1.56 | .03 | .22 |
| | Sentence Number x Token-Type | .13 | .00 | .72 | .06 | .00 | .81 |
| | Match x Sentence Number x Token-Type | 8.31 | .13 | .01 | 5.05 | .08 | .04 |

# Appendix 19: ANOVA participant analysis cell means for the pronoun region in Experiment 2

Cell means (RT) and standard deviations (SD) of the eye-movement measures for the pronoun interest area (in msec and number of regressions) as determined by the participant analyses in the ANOVA (factors: token-type (token/type), match (match/mismatch), sentence number (sen 1/sen 2); N = 40 for sen 1, N = 40 for sen 2)

| | Token | | | | Type | | | |
|---|---|---|---|---|---|---|---|---|
| | Match | | Mismatch | | Match | | Mismatch | |
| | Sen 1 | Sen 2 | Sen 1 | Sen 2 | Sen 1 | Sen 2 | Sen 1 | Sen 2 |
| Mean First Fixation Duration | 217 | 213 | 232 | 221 | 214 | 214 | 224 | 228 |
| First Fixation Duration: SD | 31 | 31 | 41 | 47 | 29 | 39 | 38 | 34 |
| Mean First-Pass Duration | 241 | 224 | 261 | 240 | 228 | 226 | 252 | 247 |
| First-Pass Duration: SD | 46 | 43 | 52 | 57 | 44 | 49 | 53 | 41 |
| Mean Selective Regression-Path Duration | 246 | 236 | 279 | 258 | 236 | 239 | 268 | 261 |
| Selective Regression-Path Duration: SD | 46 | 59 | 56 | 67 | 48 | 56 | 64 | 47 |
| Mean Total Reading Time | 309 | 262 | 355 | 322 | 298 | 298 | 354 | 333 |
| Total Reading Time: SD | 72 | 85 | 94 | 90 | 80 | 95 | 97 | 87 |
| Mean Regression Out | 0.06 | 0.09 | 0.16 | 0.13 | 0.08 | 0.12 | 0.11 | 0.09 |
| Regression Out: SD | 0.11 | 0.18 | 0.17 | 0.20 | 0.13 | 0.15 | 0.15 | 0.14 |
| Mean Regression In | 0.13 | 0.09 | 0.18 | 0.18 | 0.14 | 0.16 | 0.13 | 0.16 |
| Regression In: SD | 0.13 | 0.17 | 0.20 | 0.18 | 0.14 | 0.18 | 0.18 | 0.17 |

# Appendix 20: ANOVA results for the pronoun region in Experiment 2

| | Factors and interactions | $F_1$ (1,39) | MSE | P | $F_2$ (1,23) | MSE | P |
|---|---|---|---|---|---|---|---|
| First Fixation Duration | Match | 13.31 | 10892.31 | .00 | 5.51 | 6575.87 | .03 |
| | Sentence Number | .57 | 707.22 | .45 | .44 | 285.82 | .51 |
| | Token-Type | .08 | 70.61 | .78 | .29 | 133.63 | .59 |
| | Match x Sentence Number | .06 | 39.59 | .82 | .06 | 30.91 | .81 |
| | Match x Token-Type | .00 | .67 | .98 | .07 | 48.44 | .80 |
| | Sentence Number x Token-Type | 3.36 | 1733.34 | .08 | 1.51 | 648.64 | .23 |
| | Match x Sentence Number x Token-Type | .69 | 799.73 | .41 | .98 | 409.62 | .33 |
| First-Pass Duration | Match | 28.36 | 32448.56 | .00 | 7.72 | 17147.97 | .01 |
| | Sentence Number | 5.83 | 10305.91 | .02 | 2.75 | 3485.53 | .11 |
| | Token-Type | .47 | 893.22 | .50 | 1.38 | 1174.83 | .25 |
| | Match x Sentence Number | .14 | 273.74 | .71 | .13 | 162.69 | .72 |
| | Match x Token-Type | .24 | 317.39 | .63 | .01 | 13.41 | .92 |
| | Sentence Number x Token-Type | 4.56 | 4568.43 | .04 | 3.46 | 2302.01 | .08 |
| | Match x Sentence Number x Token-Type | .00 | 4.49 | .97 | .20 | 231.92 | .66 |
| Selective Regression-Path Duration | Match | 42.93 | 60003.27 | .00 | 13.92 | 35212.40 | .00 |
| | Sentence Number | 3.13 | 7004.28 | .09 | .63 | 1660.92 | .44 |
| | Token-Type | .55 | 1113.18 | .46 | .79 | 908.24 | .38 |
| | Match x Sentence Number | .88 | 2115.01 | .36 | .45 | 592.03 | .51 |
| | Match x Token-Type | .00 | 4.56 | .96 | .45 | 794.42 | .51 |
| | Sentence Number x Token-Type | 1.99 | 3552.58 | .17 | 2.01 | 2518.58 | .17 |
| | Match x Sentence Number x Token-Type | .00 | 6.06 | .97 | .01 | 20.40 | .92 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| Total Reading Time | Match | 54.43 | 193506.58 | .00 | 21.62 | 117592.70 | .00 |
|  | Sentence Number | 5.79 | 50715.04 | .02 | 3.32 | 25824.62 | .08 |
|  | Token-Type | 1.20 | 6386.56 | .28 | .97 | 2913.54 | .34 |
|  | Match x Sentence Number | .03 | 239.09 | .86 | .00 | 20.00 | .95 |
|  | Match x Token-Type | .12 | 1005.08 | .73 | .911 | 4702.60 | .35 |
|  | Sentence Number x Token-Type | 3.14 | 16840.31 | .08 | 2.30 | 10190.99 | .14 |
|  | Match x Sentence Number x Token-Type | 1.01 | 6102.80 | .32 | 1.00 | 3899.26 | .33 |
| Regression Out | Match | 6.11 | .10 | .02 | 5.42 | .08 | .03 |
|  | Sentence Number | .01 | .00 | .94 | .04 | .00 | .04 |
|  | Token-Type | .43 | .01 | .52 | .04 | .00 | .85 |
|  | Match x Sentence Number | 2.09 | .06 | .16 | 1.79 | .01 | .19 |
|  | Match x Token-Type | 4.21 | .08 | .05 | 5.22 | .07 | .03 |
|  | Sentence Number x Token-Type | .12 | .00 | .73 | .44 | .00 | .51 |
|  | Match x Sentence Number x Token-Type | .03 | .00 | .87 | .20 | .00 | .66 |
| Regression In | Match | 4.21 | .09 | .05 | 4.71 | .10 | .04 |
|  | Sentence Number | .00 | .00 | .99 | .23 | .01 | .64 |
|  | Token-Type | .03 | .00 | .86 | .05 | .00 | .83 |
|  | Match x Sentence Number | .46 | .01 | .50 | .56 | .01 | .46 |
|  | Match x Token-Type | 3.47 | .10 | .07 | 6.18 | .08 | .02 |
|  | Sentence Number x Token-Type | 1.85 | .04 | .18 | .70 | .02 | .41 |
|  | Match x Sentence Number x Token-Type | .23 | .01 | .64 | .66 | .01 | .43 |

# Appendix 21: ANOVA participant analysis cell means for the pronoun spill-over region in Experiment 2

Cell means (RT) and standard deviations (SD) of the eye-movement measures for the pronoun spill-over interest area (in msec and number of regressions) as determined by the participant analyses in the ANOVA (factors: token-type (token/type), match (match/mismatch), sentence number (sen 1/sen 2); N = 40 for sen 1, N = 40 for sen 2)

| | Token | | | | Type | | | |
|---|---|---|---|---|---|---|---|---|
| | Match | | Mismatch | | Match | | Mismatch | |
| | Sen 1 | Sen 2 | Sen 1 | Sen 2 | Sen 1 | Sen 2 | Sen 1 | Sen 2 |
| Mean First Fixation Duration | 232 | 223 | 244 | 236 | 223 | 216 | 235 | 217 |
| First Fixation Duration: SD | 39 | 41 | 47 | 46 | 36 | 34 | 42 | 34 |
| Mean First-Pass Duration | 276 | 259 | 308 | 282 | 271 | 268 | 283 | 268 |
| First-Pass Duration: SD | 70 | 65 | 80 | 61 | 84 | 73 | 72 | 71 |
| Mean Selective Regression-Path Duration | 301 | 269 | 337 | 305 | 287 | 306 | 313 | 308 |
| Selective Regression-Path Duration: SD | 79 | 66 | 90 | 67 | 91 | 100 | 82 | 93 |
| Mean Total Reading Time | 374 | 329 | 419 | 357 | 327 | 360 | 391 | 406 |
| Total Reading Time: SD | 111 | 100 | 144 | 92 | 130 | 129 | 122 | 180 |
| Mean Regression Out | 0.13 | 0.13 | 0.16 | 0.15 | 0.11 | 0.22 | 0.16 | 0.20 |
| Regression Out: SD | 0.18 | 0.18 | 0.18 | 0.17 | 0.18 | 0.26 | 0.22 | 0.24 |
| Mean Regression In | 0.17 | 0.20 | 0.18 | 0.14 | 0.06 | 0.12 | 0.16 | 0.19 |
| Regression In: SD | 0.18 | 0.24 | 0.21 | 0.15 | 0.13 | 0.17 | 0.22 | 0.18 |

# Appendix 22: ANOVA results for the spill-over region in Experiment 2

| | Factors and interactions | $F_1$ (1,39) | MSE | P | $F_2$ (1,23) | MSE | P |
|---|---|---|---|---|---|---|---|
| First Fixation Duration | Match | 7.34 | 7721.57 | .01 | 5.58 | 4885.58 | .03 |
| | Sentence Number | 5.31 | 9179.01 | .03 | 10.48 | 7800.45 | .00 |
| | Token-Type | 8.28 | 9440.49 | .01 | 2.83 | 1764.43 | .11 |
| | Match x Sentence Number | .26 | 363.40 | .61 | 1.50 | 924.89 | .23 |
| | Match x Token-Type | .64 | 763.76 | .43 | 1.56 | 1035.28 | .23 |
| | Sentence Number x Token-Type | .36 | 409.99 | .55 | .06 | 34.27 | .81 |
| | Match x Sentence Number x Token-Type | .61 | 723.51 | .44 | .66 | 835.84 | .43 |
| First-Pass Duration | Match | 4.56 | 22183.63 | .04 | 9.48 | 15030.31 | .01 |
| | Sentence Number | 4.25 | 18928.01 | .05 | 5.50 | 20797.72 | .03 |
| | Token-Type | 1.65 | 5816.35 | .21 | .80 | 2299.17 | .38 |
| | Match x Sentence Number | .98 | 2116.81 | .33 | 1.26 | 2121.68 | .27 |
| | Match x Token-Type | 2.71 | 9272.22 | .11 | 9.13 | 11213.23 | .01 |
| | Sentence Number x Token-Type | 1.06 | 3321.43 | .31 | 1.17 | 1416.63 | .29 |
| | Match x Sentence Number x Token-Type | .03 | 96.09 | .88 | .00 | 1.13 | .98 |
| Selective Regression-Path Duration | Match | 11.01 | 49155.10 | .00 | 8.72 | 33931.50 | .01 |
| | Sentence Number | 1.99 | 12475.51 | .17 | 3.39 | 22839.43 | .08 |
| | Token-Type | .00 | 11.85 | .96 | .12 | 516.21 | .73 |
| | Match x Sentence Number | 1.13 | 3436.39 | .30 | 1.96 | 3760.19 | .18 |
| | Match x Token-Type | 3.16 | 9539.93 | .08 | 6.98 | 14906.87 | .02 |
| | Sentence Number x Token-Type | 7.36 | 31181.30 | .01 | 6.57 | 14905.82 | .02 |
| | Match x Sentence Number x Token-Type | .55 | 2940.56 | .46 | .19 | 680.56 | .66 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Total Reading Time | Match | 15.58 | 167282.63 | .00 | 19.48 | 87758.49 | .00 |
| | Sentence Number | 1.44 | 16998.97 | .24 | 1.75 | 30234.71 | .20 |
| | Token-Type | .01 | 128.69 | .93 | .04 | 315.62 | .85 |
| | Match x Sentence Number | .78 | 5999.04 | .38 | 2.79 | 14673.59 | .11 |
| | Match x Token-Type | .68 | 7029.66 | .42 | .07 | 304.29 | .79 |
| | Sentence Number x Token-Type | 11.95 | 119952.62 | .00 | 6.00 | 56247.76 | .02 |
| | Match x Sentence Number x Token-Type | .00 | 2.71 | .99 | .04 | 141.06 | .85 |
| Regression Out | Match | .51 | .02 | .48 | .75 | .01 | .39 |
| | Sentence Number | 2.77 | .10 | .10 | .49 | .02 | .49 |
| | Token-Type | 1.43 | .05 | .24 | .48 | .01 | .50 |
| | Match x Sentence Number | .89 | .02 | .35 | .45 | .01 | .51 |
| | Match x Token-Type | .05 | .00 | .82 | 1.48 | .02 | .24 |
| | Sentence Number x Token-Type | 4.07 | .10 | .05 | 2.15 | .04 | .16 |
| | Match x Sentence Number x Token-Type | .29 | .02 | .59 | .35 | .01 | .56 |
| Regression In | Match | 3.11 | .08 | .09 | .58 | .01 | .45 |
| | Sentence Number | .98 | .04 | .33 | .87 | .04 | .36 |
| | Token-Type | 2.91 | .11 | .10 | 3.48 | .10 | .08 |
| | Match x Sentence Number | 1.53 | .04 | .22 | 3.62 | .07 | .07 |
| | Match x Token-Type | 7.55 | .25 | .01 | 2.85 | .09 | .11 |
| | Sentence Number x Token-Type | 2.18 | .06 | .15 | 3.39 | .06 | .08 |
| | Match x Sentence Number x Token-Type | .08 | .00 | .79 | .04 | .00 | .84 |

## Appendix 23: Sentence stimuli in Experiment 3

Last week, the drunken lorry driver almost killed himself/herself driving through a red light, really scared an old man on the footpath and, in addition, completely embarrassed himself/herself by not knowing the route to Cardiff.

The article stated that the goalkeeper blamed himself/herself for losing the game, decided to take a short break from the team and, in the new season, promised to devote himself/herself completely to training.

In the past, the young construction worker had often praised himself/herself for being punctual and for covering for an older workmate who frequently ran late, but lately had allowed himself/herself the luxury of being a little late as well.

Every week, the locksmith taught himself/herself a new skill using a handbook written by an American expert and, through this routine quickly established himself/herself as particularly competent.

In the evening, the young mechanic seated himself/herself comfortably in front of the TV, watched the all-night song contest with an old friend and, at bedtime, found it difficult to drag himself/herself away from the program.

The overworked security guard trusted himself/herself to do a good job but had overlooked several suspicious parcels and was criticised by the supervisor, and therefore had to acquaint himself/herself with the complicated regulations again.

In the evening, the butcher washed himself/herself thoroughly and visited the village fair with a neighbour, but did not enjoy himself/herself there and went home early.

During the journey, the experienced pilot injured himself/herself quite badly, was told by the doctor to take a long holiday and, after returning to the job, had to familiarise himself/herself with the cockpit again.

In the afternoon, the bricklayer upset himself/herself by damaging the tools and was asked by the foreman to consider further training, but decided to restrict himself/herself to less demanding jobs.

Last night, the typist introduced herself/himself to the guests at the company party as the new member of the administrative team, only hours after having excused herself/himself from another party.

One morning, the beautician spoke aloud to herself/himself about serious family problems without realising that some colleagues were listening from the next room, and was deeply ashamed of herself/himself on learning that they were gossiping about these problems.

Last week, the babysitter cut herself/himself on a piece of broken glass and almost fainted before the children’s eyes but, in spite of the injury, forced herself/himself to read to the children until the parents returned.

After a while, the florist was proud of herself/himself and liked the job in spite of the grumpy colleague working in the greenhouse, but nevertheless thought of going into business for herself/himself as soon as possible.

At times, the trainee childminder asked herself/himself whether the children’s diet was right and finally decided to consult an experienced nutritionist, having previously only set herself/himself the target of providing a bit of fresh fruit every day.

Throughout the years, the housekeeper had often criticized herself/himself for forgetting birthdays, finally asked a friend for advice and, from then on, used a calendar to remind herself/himself of important dates.

The elderly secretary thoroughly familiarised herself/himself with the new computer a few months before retiring and, to everyone’s surprise, really enjoyed herself/himself while exploring the potential of the computer.

# Appendix 24: Instructions for the memory questionnaire in Experiment 3

This is the last part of the experiment.

You will again be presented with sentences. These will be identical or slightly different from the sentences you read earlier. If they are different, then only in the agent's gender. Please read each of the sentences carefully and then decide whether you have seen it during the first part of the experiment or not. For responding with 'yes', please press the key with the 'Y' sticker, for 'no' please press the key with the 'N' sticker.

If you have any questions, please ask the experimenter now. Otherwise press the space bar to start.

Thanks!

# Appendix 25: ANOVA participant analysis cell means for the agent region in Experiment 3

Means and standard deviations (SD) of the eye-movement measures per condition for the agent area as determined by the participant analyses (N = 32 for match, N = 32 for mismatch)

| | Match | Mismatch |
|---|---|---|
| Mean First Fixation Duration | 225 | 221 |
| First Fixation Duration: SD | 41 | 38 |
| Mean First-Pass Duration | 318 | 335 |
| First-Pass Duration: SD | 71 | 76 |
| Mean Selective Regression-Path Duration | 346 | 357 |
| Selective Regression-Path Duration: SD | 74 | 83 |
| Mean Total Reading Time | 460 | 498 |
| Total Reading Time: SD | 187 | 198 |
| Mean Regression Out | 0.13 | 0.11 |
| Regression Out: SD | 0.13 | 0.17 |
| Mean Regression In | 0.18 | 0.23 |
| Regression In: SD | 0.19 | 0.24 |

## Appendix 26: ANOVA results for the agent region in Experiment 3

| | Factor | $F_1$ (1,31) | MSE | P | $F_2$ (1,15) | MSE | P |
|---|---|---|---|---|---|---|---|
| First Fixation Duration | Match | .33 | 187.55 | .57 | .02 | 6.70 | .89 |
| First-Pass Duration | Match | .91 | 4338.36 | .35 | 1.28 | 2228.78 | .28 |
| Selective Regression-Path Duration | Match | .53 | 2142.19 | .47 | .73 | 1238.03 | .41 |
| Total Reading Time | Match | 1.87 | 23171.69 | .18 | 3.16 | 12516.78 | .10 |
| Regression Out | Match | .35 | .01 | .56 | .10 | .00 | .76 |
| Regression In | Match | 3.02 | .04 | .09 | 2.56 | .01 | .13 |

# Appendix 27: ANOVA participant analysis cell means for the pronoun region in Experiment 3

Means and standard deviations (SD) of the eye-movement measures per condition for the pronoun area as determined by the participant analyses (N = 32 for pronoun 1, N = 32 for pronoun 2)

| | Match | | Mismatch | |
|---|---|---|---|---|
| | Pronoun 1 | Pronoun 2 | Pronoun 1 | Pronoun 2 |
| Mean First Fixation Duration | 216 | 203 | 223 | 201 |
| First Fixation Duration: SD | 35 | 36 | 32 | 33 |
| Mean First-Pass Duration | 231 | 215 | 247 | 216 |
| First-Pass Duration: SD | 43 | 50 | 56 | 40 |
| Mean Selective Regression-Path Duration | 236 | 220 | 277 | 220 |
| Selective Regression-Path Duration: SD | 44 | 54 | 81 | 39 |
| Mean Total Reading Time | 303 | 295 | 409 | 267 |
| Total Reading Time: SD | 83 | 109 | 153 | 71 |
| Mean Regression Out | 0.07 | 0.06 | 0.15 | 0.06 |
| Regression Out: SD | 0.13 | 0.11 | 0.17 | 0.09 |
| Mean Regression In | 0.14 | 0.13 | 0.28 | 0.14 |
| Regression In: SD | 0.13 | 0.17 | 0.21 | 0.14 |

# Appendix 28: ANOVA results for the pronoun region in Experiment 3

| | Factor | $F_1$ (1,31) | MSE | P | $F_2$ (1,15) | MSE | P |
|---|---|---|---|---|---|---|---|
| First Fixation Duration | Match | .50 | 222.76 | .49 | .02 | 6.56 | .90 |
| | Pronoun Number | 17.96 | 9207.92 | .00 | 9.01 | 4601.42 | .01 |
| | Match x Pronoun Number | 1.49 | 631.90 | .23 | 1.07 | 360.38 | .32 |
| First-Pass Duration | Match | 1.94 | 2210.29 | .17 | 1.66 | 1012.91 | .22 |
| | Pronoun Number | 14.53 | 17660.25 | .00 | 7.01 | 10696.47 | .02 |
| | Match x Pronoun Number | 2.05 | 1963.45 | .16 | 2.60 | 1722.56 | .13 |
| Selective Regression-Path Duration | Match | 6.82 | 13381.05 | .01 | 6.37 | 5859.33 | .02 |
| | Pronoun Number | 20.85 | 43100.85 | .00 | 12.73 | 23780.34 | .00 |
| | Match x Pronoun Number | 7.43 | 13035.67 | .01 | 7.23 | 7392.78 | .02 |
| Total Reading Time | Match | 7.74 | 48434.39 | .01 | 7.33 | 22459.14 | .02 |
| | Pronoun Number | 29.52 | 180400.72 | .00 | 12.81 | 88636.45 | .00 |
| | Match x Pronoun Number | 25.26 | 143314.53 | .00 | 34.19 | 65458.58 | .00 |
| Regression Out | Match | 2.54 | .06 | .12 | 5.16 | .04 | .04 |
| | Pronoun Number | 425 | .08 | .05 | 5.91 | .06 | .03 |
| | Match x Pronoun Number | 3.57 | .04 | .07 | 4.34 | .03 | .06 |
| Regression In | Match | 9.24 | .20 | .01 | 4.97 | .10 | .04 |
| | Pronoun Number | 9.17 | .17 | .01 | 4.30 | .07 | .06 |
| | Match x Pronoun Number | 6.17 | .14 | .02 | 7.25 | .05 | .02 |

# Appendix 29: ANOVA participant analysis cell means for the pronoun spill-over region in Experiment 3

Means and standard deviations (SD) of the eye-movement measures per condition for the pronoun spill-over area as determined by the participant analyses (N = 32 for pronoun 1, N = 32 for pronoun 2)

| | Match | | Mismatch | |
|---|---|---|---|---|
| | Pronoun 1 | Pronoun 2 | Pronoun 1 | Pronoun 2 |
| Mean First Fixation Duration | 228 | 220 | 232 | 218 |
| First Fixation Duration: SD | 48 | 40 | 43 | 34 |
| Mean First-Pass Duration | 280 | 277 | 296 | 277 |
| First-Pass Duration: SD | 59 | 61 | 59 | 71 |
| Mean Selective Regression-Path Duration | 300 | 310 | 342 | 307 |
| Selective Regression-Path Duration: SD | 65 | 77 | 94 | 119 |
| Mean Total Reading Time | 399 | 388 | 427 | 400 |
| Total Reading Time: SD | 134 | 127 | 128 | 142 |
| Mean Regression Out | 0.11 | 0.14 | 0.23 | 0.14 |
| Regression Out: SD | 0.11 | 0.14 | 0.17 | 0.16 |
| Mean Regression In | 0.18 | 0.16 | 0.14 | 0.16 |
| Regression In: SD | 0.17 | 0.14 | 0.14 | 0.14 |

# Appendix 30: ANOVA results for the spill-over region in Experiment 3

| | Factor | $F_1$ (1,31) | MSE | P | $F_2$ (1,15) | MSE | P |
|---|---|---|---|---|---|---|---|
| First Fixation Duration | Match | .03 | 29.15 | .86 | .17 | 74.93 | .69 |
| | Pronoun Number | 3.08 | 3678.68 | .09 | 3.20 | 1715.31 | .09 |
| | Match x Pronoun Number | .40 | 294.33 | .53 | 1.23 | 536.68 | .28 |
| First-Pass Duration | Match | .95 | 2319.15 | .34 | .81 | 1294.29 | .38 |
| | Pronoun Number | .93 | 4065.54 | .34 | 2.33 | 8474.35 | .15 |
| | Match x Pronoun Number | .75 | 2064.35 | .39 | 1.70 | 2438.26 | .21 |
| Selective Regression-Path Duration | Match | 2.47 | 12702.39 | .13 | 2.50 | 6767.94 | .13 |
| | Pronoun Number | .67 | 4717.24 | .42 | 2.34 | 16856.48 | .15 |
| | Match x Pronoun Number | 2.07 | 15925.99 | .16 | 3.94 | 11535.83 | .07 |
| Total Reading Time | Match | 1.34 | 13071.22 | .26 | 2.80 | 7210.13 | .12 |
| | Pronoun Number | 1.49 | 11919.10 | .23 | 1.82 | 32956.77 | .20 |
| | Match x Pronoun Number | .23 | 2284.96 | .63 | .90 | 4159.93 | .36 |
| Regression Out | Match | 5.48 | .11 | .03 | 4.40 | .06 | .05 |
| | Pronoun Number | 1.81 | .03 | .19 | .35 | .01 | .56 |
| | Match x Pronoun Number | 6.93 | .12 | .01 | 3.09 | .04 | .10 |
| Regression In | Match | .94 | .01 | .34 | 1.58 | .01 | .23 |
| | Pronoun Number | .01 | .00 | .91 | .21 | .00 | .65 |
| | Match x Pronoun Number | .36 | .01 | .55 | .30 | .00 | .59 |

## Appendix 31: Sentence stimuli in Experiment 4 to 6: pronoun condition (1) and no-pronoun condition (2)

(1) On Monday the babysitter cut herself on a piece of broken glass.

(2) On Monday the babysitter came across a piece of broken glass.

(1) At times the beautician spoke to herself when working alone.

(2) At times the beautician sang aloud when working alone.

(1) In the afternoon the bricklayer upset himself by damaging the tools.

(2) In the afternoon the bricklayer lost time by damaging the tools.

(1) In the evening the butcher washed himself thoroughly and went out.

(2) In the evening the butcher washed the dishes and went out.

(1) In the end the carpenter convinced himself that the material was faulty.

(2) In the end the carpenter was convinced that the material was faulty.

(1) On Saturday the cheerleader dressed herself in a bright costume.

(2) On Saturday the cheerleader was dressed in a bright costume.

(1) At times the childminder asked herself if the children’s diet was right.

(2) At times the childminder was asked if the children’s diet was right.

(1) Quite often the construction worker praised himself for being punctual.

(2) Quite often the construction worker was praised for being punctual.

(1) After a while the florist was proud of herself and really liked the job.

(2) After a while the florist became more skilled and really liked the job.

(1) On a Sunday the fortune teller treated herself to cakes with cream.

(2) On a Sunday the fortune teller was treated to cakes with cream.

(1) The article stated that the footballer blamed himself for losing the game.

(2) The article stated that the goalkeeper was blamed for losing the game.

(1) Many times the housekeeper criticised herself for forgetting birthdays.

(2) Many times the housekeeper was criticised for forgetting birthdays.

(1) Every week the locksmith taught himself another little skill.

(2) Every week the locksmith acquired another little skill.

(1) Last week the lorry driver almost killed himself by driving without lights on.
(2) Last week the lorry driver almost caused an accident by driving without lights on.

(1) In the evening the mechanic seated himself comfortably in front of the TV.
(2) In the evening the mechanic sat down comfortably in front of the TV.

(1) A month ago the midwife bought herself a new working uniform.
(2) A month ago the midwife bought a new working uniform.

(1) At weekends the nanny was comfortable with herself in the large house.
(2) At weekends the nanny felt very comfortable in the large house.

(1) During the journey the pilot injured himself quite badly.
(2) During the journey the pilot was injured quite badly.

(1) After work the plumber got himself a big portion of chips.
(2) After work the plumber got a big portion of chips.

(1) On several occasions the receptionist hurt herself with the sharp scissors.
(2) On several occasions the receptionist cut the flowers with the sharp scissors.

(1) Last week the secretary familiarised herself with the new photocopier.
(2) Last week the secretary became familiarised with the new photocopier.

(1) Most of the time the security guard trusted himself to do a good job.
(2) Most of the time the security guard was trusted to do a good job.

(1) Often during the day the taxi driver looked at himself in the rear view mirror.
(2) Often during the day the taxi driver looked at the traffic in the rear view mirror.

(1) Last night the typist introduced herself to the other party guests.
(2) Last night the typist was introduced to the other party guests.

## Appendix 32: Instructions for Experiment 4

In this task you will read a number of sentences and answer questions about them. We would like you to read these sentences carefully so you can answer the questions correctly! Once you have finished reading a sentence, please press the Y or N key. If you want to respond to a question with NO, please press the N key, if you want to respond with YES, please press the Y key. To make the task more challenging, you will additionally see photos of women and men. Your task here is simply to report the individuals' gender -- as QUICKLY and ACCURATELY as you can. Please press the F key for FEMALE, and the M key for MALE. Again: you will read sentences. After each sentence you either respond to a question OR categorise a picture. Please position your fingers on the response keys and press the spacebar when you are ready to begin.

## Appendix 33: Instructions for Experiment 5

In this task you will read a number of sentences and answer questions about them. We would like you to read these sentences carefully so you can answer the questions correctly! Once you have finished reading a sentence, please press the spacebar. If you want to respond to a question with NO, please press the N key, if you want to respond with YES, please press the Y key. To make the task more challenging, you will additionally see a photo between the sentence and the question. Your task here is simply to report the individuals' gender. It is important that you do this QUICKLY and ACCURATELY. Please press the F key for FEMALE, and the M key for MALE. At the beginning of each trial you will see a fixation cross. Again: on each trial you will see a fixation cross, read a sentence, then categorise a picture and after that respond to a question about the sentence. Before the main task, however, we will give you some training with categorising the pictures. Please position your fingers on the response keys and press the spacebar when you are ready to begin.

## Appendix 34: Instructions for Experiment 6

Welcome and thanks for taking part in this study! Please use the spacebar to go through the instructions. The first two blocks will consist of a face recognition task. In the remaining three blocks you will have to switch between the face recognition task and a simple reading task. Your task in the face recognition task is to simply report the individuals' gender. It is important that you do this as QUICKLY and ACCURATELY as you can! Please press the F key for FEMALE, and the M key for MALE. In the sentence reading task, please read the sentences as you would read a newspaper article. Here you are under no time pressure. Once you have finished reading a sentence, please press the spacebar. For now let's start with the blocks with the face recognition task only. Please position your fingers on the response keys and press the spacebar when you are ready to begin.

## Appendix 35: Instructions for Experiment 7

Welcome and thanks for taking part in this study! Please use the spacebar to go through the instructions. The first two blocks will consist of a face recognition task. Your task here is to simply report the individuals' gender. It is important that you do this as QUICKLY and ACCURATELY as you can! Please press the F key for FEMALE, and the M key for MALE. In the next three blocks a word will be presented shortly before the picture. Please just read that word and carry on with the face recognition task. For now let's start with the blocks with the face recognition task only. Please position your fingers on the response keys and press the spacebar when you are ready to begin.

# Appendix 36: ANOVA results for Experiments 4 - 7

| Experiment 4 | Factors and interactions | $F_1$ (1,27) | MSE | P | $F_2$ (1,20) | MSE | P |
|---|---|---|---|---|---|---|---|
| | Match | .01 | 105.99 | .91 | .01 | 48.40 | .93 |
| | Picture | 3.29 | 34221.87 | .08 | 1.44 | 8544.57 | .24 |
| | Sentence | 5.79 | 45761.15 | .02 | 2.31 | 11803.44 | .15 |
| | Match x Picture | 5.24 | 58986.96 | .03 | 1.68 | 9967.15 | .21 |
| | Match x Sentence | 3.21 | 29638.72 | .08 | 1.51 | 7734.54 | .23 |
| | Picture x Sentence | .54 | 6070.07 | .47 | .35 | 1791.79 | .56 |
| | Match x Picture x Sentence | 7.70 | 65952.74 | .01 | 2.40 | 12288.70 | .14 |
| Experiment 5 | Factors and interactions | $F_1$ (1,19) | MSE | P | $F_2$ (1,20) | MSE | P |
| | Match | .24 | 1691.11 | .63 | .00 | 3.66 | .97 |
| | Picture | 1.50 | 9484.09 | .24 | 2.44 | 5518.30 | .13 |
| | Sentence | .86 | 6753.63 | .36 | .31 | 729.07 | .59 |
| | Match x Picture | .20 | 1104.97 | .66 | .00 | .62 | .99 |
| | Match x Sentence | 4.02 | 39767.84 | .06 | 4.02 | 9571.19 | .06 |
| | Picture x Sentence | 3.12 | 14768.07 | .09 | 1.03 | 2465.06 | .32 |
| | Match x Picture x Sentence | .09 | 878.39 | .77 | .00 | .56 | .99 |
| Experiment 6 | Factors and interactions | $F_1$ (1,19) | MSE | P | $F_2$ (1,20) | MSE | P |
| | Match | 1.20 | 6489.25 | .29 | 1.08 | 1720.04 | .31 |
| | Picture | 1.00 | 5198.17 | .33 | .84 | 1333.61 | .37 |
| | Sentence | 1.22 | 7457.54 | .28 | .81 | 2106.79 | .38 |
| | Match x Picture | 5.46 | 25554.54 | .03 | 5.04 | 8010.56 | .04 |
| | Match x Sentence | .67 | 5597.25 | .43 | .69 | 1796.02 | .42 |
| | Picture x Sentence | 2.86 | 10758.40 | .11 | 1.34 | 347098 | .26 |
| | Match x Picture x Sentence | .70 | 4666.03 | .41 | .74 | 1926.85 | .40 |

| Experiment 7 | Factors and interactions | $F_1$ (1,11) | MSE | P | $F_2$ (1,20) | MSE | P |
|---|---|---|---|---|---|---|---|
| | Match | 19.66 | 42237.62 | .00 | 20.95 | 18224.33 | .00 |
| | Picture | .35 | 510.84 | .57 | .04 | 34.43 | .84 |
| | Match x Picture | 1.48 | 2429.77 | .25 | 1.04 | 906.72 | .32 |