#NotAllStereotypes: Examining the Effect of Speaker Inconsistency on Reading Comprehension

Gareth Dwyer

Saarland University

Author Note:

This research was conducted in collaboration with Dr. Les Sikos and the students in the module *Experimental Methods in Psycholinguistic Research*, Saarland University, Summer, 2016

Abstract

When is speaker context taken into account? Recent studies have suggested that a listener may take the speaker's identity into account when interpreting an utterance much earlier than previously assumed. We closely follow the methods used by Van Berkum *et al.*, adapting that study to a self-paced reading one. We had 52 participants read brief fictional descriptions of speakers with some context and then timed their reading of a target sentence, uttered by the described speaker. We manipulated the speaker and context descriptions to make the target sentences either stereotypically consistent or stereotypically inconsistent with the target sentence. Although our results were not significant, we did see some indication that inconsistent sentences take longer to comprehend than consistent ones. We focus our discussion on our design and procedure, and pave the way for future work in this area. We conclude that self-paced reading remains an interesting paradigm to explore the so-called "speaker inconsistency" effect, and detail some areas in which our design and methods could be improved.

Keywords: Speaker Inconsistency, Stereotype, Self-paced Reading, comprehension

#NotAllStereotypes: The Effect of Speaker Inconsistency on Reading Comprehension

Introduction

How we process language is not fully understood. Specifically, how and when the human brains interprets the *pragmatics* of a sentence (the meaning including all context such as who the speaker is) as opposed to the *semantics* of the sentence (the literal meaning of the sentence without taking context into account), is a controversial topic. Traditionally, the interpretation process has been modeled as a multi-stage fallback process, in which the semantics of a sentence are comprehended first and pragmatics are only taken into account if the purely semantic interpretation seems unlikely or problematic. More modern models have suggested that the traditional models are not accurate and that actually we comprehend pragmatic content, including figurative language and speaker context, immediately on hearing or reading a sentence (Glucksberg, 2001; Van Berkum *et al.*, 2008).

In the current study, we closely follow the ideas and methods used by Van Berkum *et al.* (2008), who played participants recordings of speakers and manipulated whether or not the utterance was stereotypically consistent with the speaker. For example, the utterance "I have a large tattoo on my back" spoken with an upper-class accent is defined as an inconsistent utterance. The researchers found a strong N400 effect for inconsistent sentences, suggesting that the brain takes speaker context into account immediately on hearing an utterance. Where that study used recorded utterances and measured participants neural responses using Electroencephalography (EEG), we instead gave participants a self-paced reading task and measured their response times, comparing responses to consistent and inconsistent utterances.

We hypothesized that inconsistent sentences would take longer to process. If the brain does immediately take into account the speaker context, as suggested above, then an utterance that breaks a social stereotype and surprises the reader should take longer to process than an utterance that is completely consistent with social stereotypes. We created text stimuli which

each briefly described a speaker and a context and then presented a target sentence, uttered by the fictional speaker. We manipulated whether or not the target sentence was stereotypically consistent with speaker and the context. We used a 2x2 design, where the speaker description and the context description could each appear in either a positive (in line with social stereotypes) or negative (against social stereotypes) condition, resulting in four possible instantiations of each item. The independent variables were the conditions of the speaker and the context.

We defined a *critical word* for each target sentence (the word that was likely to illicit surprise in the negative stereotypical conditions), and our dependent variable was the reading time of these critical words (the target sentence was presented as a self-paced reading task) in positive and negative conditions.

We presented the items to our participants as a self-paced reading task using web-based software. We took timings of reading times for the critical word, and also for the two words which immediately followed the critical word to account for spillover effect (discussed in *Results*).

Method

Participants

Fifty-two people (27 male, 25 female), all of whom identified German as their dominant language, participated in the study. All were over 18 (average age 24), and each signed a consent form[1] agreeing to the use of their (anonymised) data. Each participant received 10 Euro for their participation.

Materials

Our stimuli design was based strongly on Van Berkum *et al.* We created stimuli in two categories: *experimental* items and *filler* items. Both sets of items contained three components. The first component is a description of a *speaker,* which gives only factual information about a

[1] A copy of this is available in the github repository here https://github.com/sixhobbits/not-all-stereotypes

fictional character (for example, "Linda is a young German woman"). The second component is a description of a *context*, which gives further details about that person's life (for example, "She just got married and bought a new home with her husband"). The third component is the *target sentence*, which is an utterance of the speaker (for example, "I want to have kids before I turn 30"). Each filler items contains a single instance of each of the above-described components, while the experimental items have alternate forms for both the speaker and the context components (two of each), resulting in each experimental item having a total of five components. The alternate forms of the speaker and context components are designed such that one describes a speaker who is stereotypically *consistent* with the target sentence while the other describes a speaker who is stereotypically *inconsistent*. An example of an experimental item with all of its components can be seen in Table 1. The 30-year-old Anna (pS) is stereotypically consistent with getting married, while it is surprising for the 75-year-old Anna (mS) to get married. The positive context, pK, (Anna has recently fallen in love) provides context that allows the target sentence to once again seem stereotypically consistent, while the negative context, mK, (Anna wants to go to China) means that the target sentence remains surprising given the 75-year-old Anna.

pS	mS	pK	mK	Target sentence
Anna is a 30 year old woman	Anna is a 75 year old woman	She has never been in serious relationships. But then she fell passionately in love with the man of her dreams last year.	She has never been to China, but then she suddenly decided to book a trip to Beijing before it was too late.	I am eagerly waiting for my **wedding** next month.

Table 1: *An example experimental item with plus Speaker (pS), minus Speaker (mS), plus Context (pK) and minus Context (mK) components.* [2]

Each experimental item has four possible forms, created from any combination of the two speaker forms and the two context forms (2x2 design). Intuitively, the mS.mK form ("Anna is a 75-year-old woman. She has never been to China, but then she suddenly decided to book a trip to

[2] All items can be seen in the project github repository (refer to previous footnote).

Beijing before it was too late") should make the target sentence the most surprising, and could therefore take longer to for the reader to comprehend.

For each target sentence, we defined a critical word (CW). The CW is the first word that is likely to seem surprising given the mS.mK condition of the item ("wedding" in the example above). We designed the target sentences such that the CW was always a common noun and never appeared at the end of the target sentence.

Stereotypes are by nature highly subjective. In order to obtain more balanced stimuli, the items were written by 13 experimenters, all of whom participated in designing and running this study. Each experimenter created five experimental items and five filler items. Each experimenter then assigned a rating (1-5) for each item, indicating how good the items were in terms of the stereotypes that they invoked. We chose the top 40 highest-rated experimental items and the top 60 filler items to be used in the final experiment.

All items were created in English and translated into German. We ensured that the critical words were common by using CELEX[3] to look at the frequencies of the words we used as well as their lemmas. We also excluded critical words that were not nouns, were compounds, or were too long.

Procedure

We carried out a self-paced reading experiment using the web-browser based software *Ibex*[4]. This software facilitates self-paced reading using JavaScript, thus allowing us to run the experiment directly on the experimenters' personal computers. Each participant was given 100 items to read, where the speaker and context components of each item were displayed as normal text, and the target sentence component was then displayed word-by-word, requiring the participant to hit the space bar in order to see the next word. Each participant saw 40

[3] http://celex.mpi.nl/
[4] http://spellout.net/ibexfarm

experimental items and 60 filler items. Additionally, the experiment started by presenting the participant with several practice items, which were discarded from the results set. We used *Ibex's* 'shuffle' and 'sequence' functionality to pseudorandomize the order of the items and the fillers, and to ensure that each participant saw each item in only one of the four conditions (mS.mK, mS.pK, pS.mK, pS.pK). On average, each participant took 40 minutes of screen time and 70 minutes total experiment time.

We asked participants about their proficiency in other languages, and each participant also completed an exit survey, in which we asked about various factors, including about previous experience with psycholinguistic experiments. Some points of interest from the exit survey are highlighted in the Discussion section. Although the 42 participants were divided amongst the 13 researchers, we ensured consistency in experience by sticking closely to a detailed experimenter script, and ensuring that all participants were run in similar environments in terms of noise levels and comfort.

Results

We did not find statistically significant results to support our hypothesis that sentences which broke common social stereotypes took longer to comprehend than sentences which were stereotypically consistent. The timings of the reading for the critical word were almost identical across all conditions. However, taking into account the so-called *spillover effect* (Vasishth, S., 2006) in which the extra mental processing caused by a word that is difficult to comprehend is in fact only observed in the words that immediately follow the problematic word, the effects of the 'surprising' critical word may be observed in the reading times of the subsequent words. Figure 1 shows the response times (RTs) for each of the four conditions for CW+1 (the word after the critical word) and CW+2 (two words after the critical word). If our hypothesis were correct, the mS.mK condition would be the most surprising and would therefore take longer to

process. Although not statistically significant, we can see that response times were slightly longer for mS.mK for CW+1 and that this difference is more pronounced in CW+2.

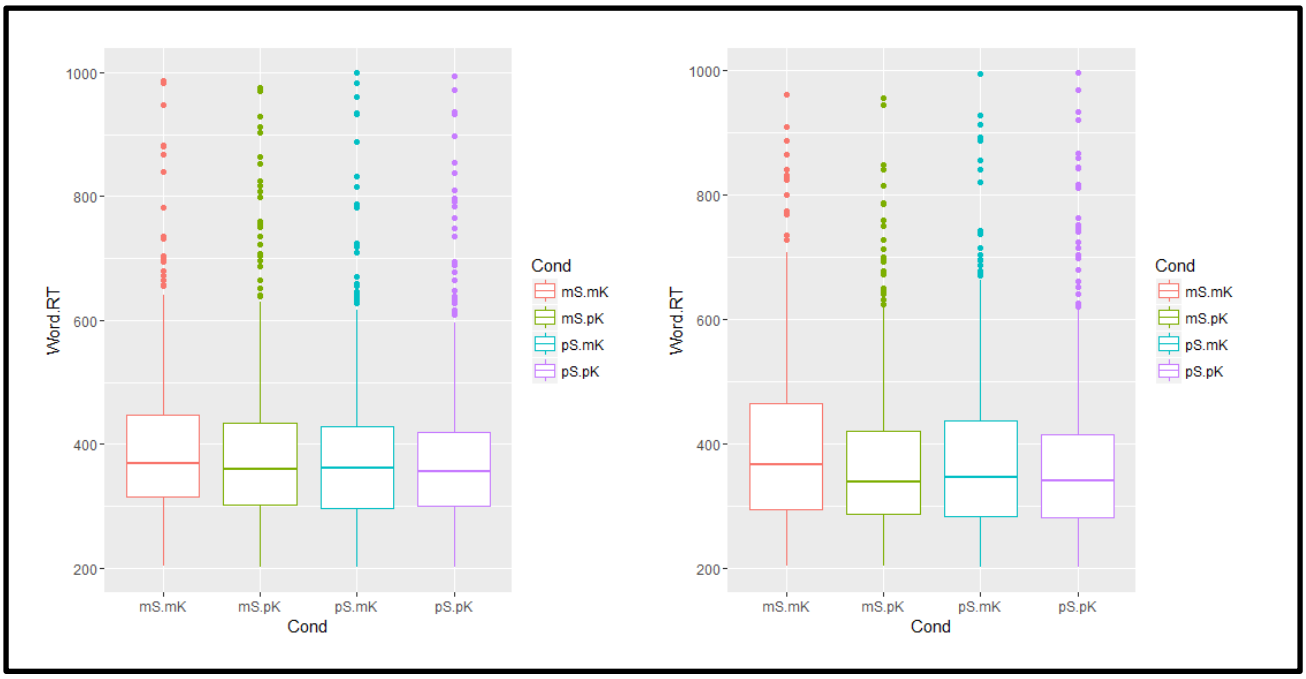

Figure 1: *Response times by condition for CW+1 (left) and CW+2 (right).*

We used an ANOVA test to analyze the different effects produced by manipulating only the speaker component, only the context component, or both speaker and context components. The combination of speaker and context produced the biggest effect (p=0.6), but again, no results were statistically significant. The ANOVA results for each of the components can be seen in Table 2.

Effect	**F**	**p**	**ges**
Speaker	0.036	0.85	2.771855e-05
Context	1.76836713	0.1896192	8.774174e-04
Speaker + Context	0.23003173	0.6335916	1.482181e-04

Table 2: *ANOVA results for Speaker and Context manipulations*

Discussion

Self-paced reading is a far less precise measuring method than EEG. While Van Berkum *et al.* saw extremely clear-cut and prominent results of an N400 effect, our current study failed to

produced similar results using self-paced reading. We therefore focus the current discussion on our methods and potential future improvements, rather than the results themselves. We first discuss some of the challenges presented by using a self-paced reading paradigm. We then present an analysis on some of the design choices we made, and the potential merits of alternatives. Finally, we look at some points that were raised by participants in the exit survey, and talk about what these might mean for the study.

Using the EEG measurement method, one can examine exactly at what time point an effect was observed in mental activity. With self-paced reading, this is more difficult, as mental activity might not immediately cause any measurable action. We partly compensated for this by examining the words that follow the critical word, as discussed above. However this is not a perfect solution. Different people seem to 'buffer' mental processing in different ways – for example, we noted some participants would press the space bar key very regularly, progressing word by word through the target sentence without a change in rhythm. Only on reading the last word, would these participants pause while they thought about and interpreted the sentence in its entirety. Furthermore, reading a sentence word by word is not how people naturally read (eye tracking studies have shown that instead readers will often backtrack and re-read some previous words if they have comprehension problems at any point while reading the sentence). Therefore, it is possible that in this artificial self-paced reading environment, participants process stereotypical understanding in a different way. Therefore, self-paced reading tasks present some challenges to researchers that are not present in, for example, EEG tasks. Nonetheless, self-paced reading is much less invasive to participants, and it remains a commonly used and useful method for gathering data about how human comprehension should be modeled (Koornneef and Van Berkum, 2006; Ferreira and Henderson, 1990).

We believe, therefore, that self-paced reading remains a worthwhile method to examine the speaker inconsistency effect. However, our design and materials have room for improvement,

and implementing such improvements could lead more clear-cut results. Perhaps our largest deviation from Van Berkum *et al.* was the omission of materials containing a *semantic anomaly.* In addition to the stereotypically inconsistent utterances that that study used, the researchers also measured the effect of sentences that were anomalous regardless of speaker (for example, "Dutch trains are *sour* and yellow" as opposed to "Dutch trains are *blue* and yellow). As anomalies of this nature are almost certain to slow down comprehension time, it would be interesting to have included similar items in our current study. We could then have compared the ely fect of the stereotypical inconsistencies with these semantic inconsistencies. We also diverged from Van Berkum *et al.* by attempting to design items that revolved around broad stereotypes, recognizable by a diverse and international group of people (our 13 experimenters were representative of several countries across Europe, Asia and Africa). Van Berkum *et al*. on the other hand, focused specifically on stereotypes that were likely to be held by Dutch people. Although it is likely that the ideas presented by Van Berkum *et al.* are generalizable, it is possible that our items were too broad in nature, and that some of the participants (many of whom were also from diverse backgrounds, although all identified as German native speakers) did not find some of the inconsistent items as surprising as we intended them to.

From the exit survey that our participants completed, we noticed another potential flaw in our design. After each item, the participants were given a comprehension question either about the speaker description or about the context, but the question was never about the target sentence itself. The more astute of are participants noticed this fact fairly quickly, and paid less attention to the target sentence as a result. Furthermore, while most participants claimed at the end that they had no idea what the study was about, several inferred that we were testing the effects of stereotypes. To counteract this, future studies might include more filler items as well as subtler inconsistencies in stereotypes (although the latter might introduce new problems of its own). Finally, half of our 52 participants had previously participated in other psycholinguistics

experiments, and nearly all of them were involved in academia in some way. It is likely that our participants therefore were not representative of the general population, and we would especially assume that older, less-educated participants would hold stronger prejudices and more stereotypical views. It would be interesting to see if a stronger effect could be found using a different set of participants.

In conclusion, although we did not find statistically significant evidence to support the claims of Van Berkum *et al.*, we presented the materials and design for a self-paced reading experiment to measure the effect of speaker inconsistency on reading comprehension. We talked about how our experiment could be modified to hopefully achieve better results in the future, and we hope that the current study can serve as a baseline to anyone who wants to use self-paced reading to examine comprehension effects.

References

Ferreira, F, & Henderson. J. M. (1990)> Use of verb information in syntactic parsing: evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning Mmemory, and Cognition, 16(4), 555*

Glucksberg, S. (2001). Understanding figurative language: From metaphor to idioms. Oxford University Press.

Koornneef, A. W., & Van Berkum, J. J. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, 54(4), 445-465.

Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. Journal of cognitive neuroscience, 20(4), 580-591.

Vasishth, S. (2006). On the proper treatment of spillover in real-time reading studies: Consequences for psycholinguistic theories. In Proceedings of the International Conference on Linguistic Evidence.