Table of Contents

1. Introduction

This is a Data Sharing Collection (DSC) from a combined eye tracking and fMRI study where 40 participants read 2 Dutch literary short stories each. The DSC includes (1) questionnaire output, (2) raw MRI data, (3) raw eye tracking data, and (4) files necessary for the interpretation and analysis of these data.

The data set was collected in the context of thesis work by Marloes Mak, within a research project of Roel Willems, entitled *'A girl fallen from the sky': Effects of sensori-motor and emotional simulation on aesthetic appreciation of literary narratives*. The data set was collected for the purpose of a study about the relationships between mental simulation, brain activity and eye movement behavior.

This codebook explains all folders and files included in this data set.

1.1 People and affiliations

Dr. Marloes Mak

- Radboud University, Nijmegen, The Netherlands

Dr. Myrthe Faber

- Radboud University, Nijmegen, The Netherlands
- Tilburg University, Tilburg, The Netherlands

Dr. Roel M. Willems

1.2 Funding

The collection of this dataset was funded by NWO (Netherlands Organization for Scientific Research; Grant number Vidi-276-89-007; granted to Roel Willems).

1.3 Publications

Publications based on this data set are

2. Methods

2.1 Pretest

In a pretest described in detail in Mak and Willems (2019), all words in the two stories used in the current experiment were rated by 30 participants on whether these words were part of a motor description, a perceptual description, or a mental event description. A total of 90 participants took part in the pretest: each type of description was rated by a different group of 30 participants. Motor descriptions are defined as "concrete acts or actions performed by a person or object," such as *"They reached the bus-stop shelter"* (Story B: Symbols & Signs). Perceptual descriptions are "things that are perceivable with the senses," such as *"A tiny unfledged bird"* (Story B: Symbols & Signs). Mental event descriptions are "explicit descriptions of the thoughts, feelings and opinions of a character" and/or "reflection[s] by a character on his own or someone else's thoughts, feelings or behaviour". For example, *"She thought of the recurrent waves of pain"* (Story B: Symbols & Signs).

It was counted how many participants underlined each word for each of the three types of description. This resulted in scores ranging from 0-30 per word, per type of description. These scores were taken as regressors for the fMRI and eye-tracking data analyses. The number of descriptive words per story per type of description can be seen in Fig. 1. There was no clear association between the ratings per word for motor, perceptual and mental event descriptions on the one hand, and other word characteristics (i.e., lexical frequency, word length, surprisal; see Appendix A). The rationale for using these ratings instead of, for example, existing ratings of concreteness (which is highly correlated with imageability; Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014) or sensory modality (Speed & Brysbaert, 2022), is that the ratings obtained in our pretest take the context of the stories into account. Since we were interested in mental simulation during the reading of complete literary short stories (a contextualized process; see Willems & Peelen, 2021), we believed it fitting to use contextualized ratings.

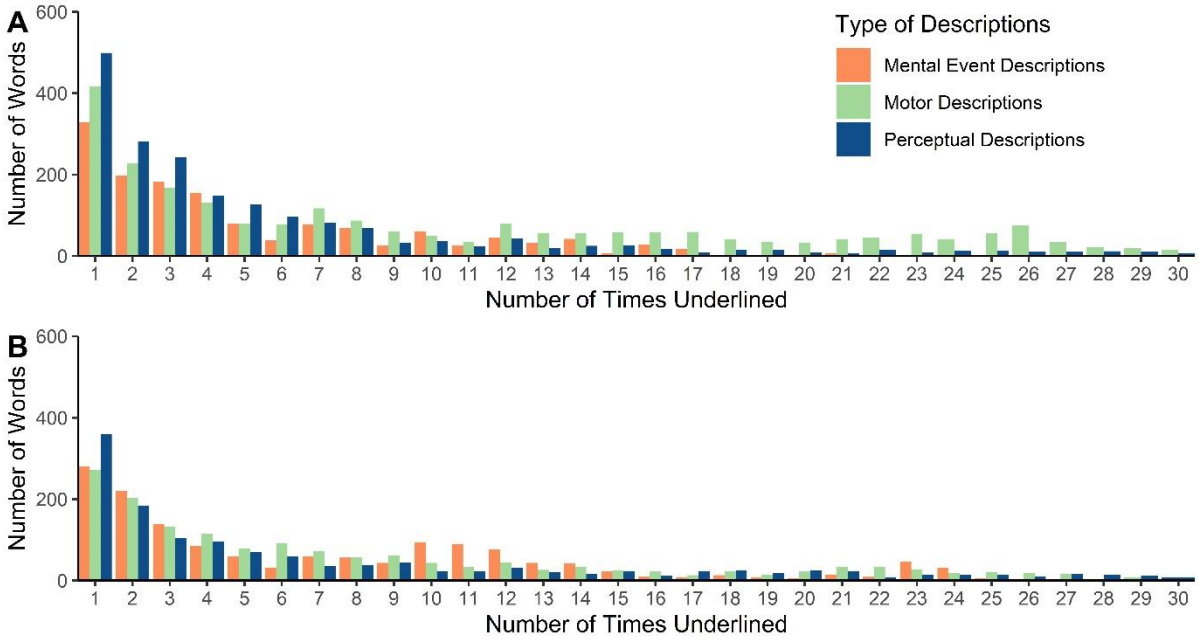

Figure 1. Number of times words were underlined for mental event, motor and perceptual descriptions in stories (A) *The people who had everything delivered* and (B) *Symbols and Signs*. Note: in story A, 1562 words, 659 words and 1070 words were underlined by none of the pre-test participants for mental event, motor and perceptual descriptions, respectively. In story B, 646, 586 and 788 words were underlined by none of the pre-test participants for mental event, motor and perceptual descriptions, respectively.

2.2 Participants

Forty participants took part in the current experiment (16 male). Participants were between 18 and 43 years old (M = 24.61, SD = 5.22). A power analysis based on Mak & Willems (2019) showed that power would be above .8 to capture small effects, in a study with 40 participants, reading two stories, with minimally 400 descriptive words per story per type of description (power analysis based on Jobe, 2009). We tested participants from the participant database of the Radboud University. Participants had no dyslexia, and had normal vision or vision correction of maximally +4 or -4 (vision correction in the scanner was done with contact lenses or MR compatible glasses that were attached to the head coil). Other exclusion criteria were epilepsy, claustrophobia, pregnancy, brain surgery, or non-removable metal in or on the body. All inclusion and exclusion criteria were established prior to data collection. Participants gave informed consent prior to the study and were allowed to withdraw their consent at any point throughout the experiment, in accordance with the declaration of Helsinki. This experiment was approved under the ethical approval of the ethical committee CMO Arnhem/Nijmegen (CMO 2014/288; version 2).

2.3 Materials

2.3.1 Stories

Two existing Dutch short stories were presented to the participants (also used in Mak & Willems, 2019). Story selection was based on the length of the stories, the presence of descriptive content, and the probability that the stories were unknown to the participants (in the study of Mak & Willems, none of the participants reported having read one of these stories before). Both stories are written by acclaimed writers and have been published by literary publishers. One story (*De mensen die alles lieten bezorgen [The people who had everything delivered]*; henceforth Story A) is written by the contemporary Dutch writer Rob van Essen (2014), and the other story (*Signalen en Symbolen*; henceforth Story B) is a professional and published translation (American English to Dutch) of *Symbols and Signs* by Vladimir Nabokov (translation in: Nabokov, 2003). The stories are 2988 and 2143 words long, respectively, and take around 10 minutes to read (Story A: *M* = 10.08, *SD* = 3.01; Story B: *M* = 9.70, *SD* = 2.94). All participants read both stories, in counterbalanced order.

2.3.2 Questionnaires

After reading the stories, participants completed a set of questionnaires. The questionnaires measuring reading experiences were filled out twice, directly after reading each story. The rest of the questionnaires were read at the end of the experiment (see also *Procedure* below).

2.3.2.1 Reading experiences

Reading experiences (i.e., story world absorption, story appreciation) were measured using questionnaires. Story world absorption was measured with the Story World Absorption Scale (SWAS; Kuijpers, Hakemulder, Tan, & Doicaru, 2014; e.g., *When I finished the story I was surprised to see that time had gone by so fast; I could imagine what the world in which the story took place looked like*), complemented with six additional questions (partly based on items originally designed by Kuijpers et al., 2014) more specifically aimed at measuring the experience of different kinds of simulation (mainly perceptual and motor simulation, e.g., *I could see the events in the story happening as if I could see through the eyes of the main character; I could easily depicture the characters in the story*). The SWAS is a validated scale consisting of 18 items with high internal validity (Kuijpers et al., 2014), which measures 4 dimensions of story world absorption via the subscales Attention, Transportation, Emotional Engagement and Mental Imagery. Participants rate each question on a 7-point scale (1 = disagree, 7 = agree). Story appreciation was measured with a questionnaire consisting of a general score of story liking (*How did you like the story*; 1 = It was very bad, 7 = It was very good) and thirteen adjectives (e.g., *[did you find the story] Entertaining, … Ominous*) that can be used to describe the stories (adapted from Knoop, Wagner, Jacobsen, & Menninghaus, 2016). These adjectives are taken from a list of adjectives that were found to be most often used by people to describe their opinion of poetry, and which can

also be used to describe aesthetic appeal in the domain of literature (Knoop et al., 2016; Mak, Faber, & Willems, under review). Finally, 6 questions are asked regarding the enjoyment of the story (from Kuijpers et al., 2014; e.g., *I was constantly curious about how the story would end; I thought the story was written well*). Participants rate both the adjectives and the questions regarding enjoyment on a 7-point scale (1 = disagree, 7 = agree). Both of these questionnaires were also used in the previous eye-tracking experiment.

2.3.2.2 Comprehension check

Story comprehension was measured using a comprehension check, consisting of 3 multiple choice questions per story with 4 possible answers per question, that should have been possible to answer correctly for people who read the stories with normal attention (example question, *Why did Jeffrey and Rita leave the flat?*). Additionally, participants were asked to indicate whether they have read any of the stories before.

2.3.2.3 Reading habits

Reading habits were assessed both directly and indirectly. The direct measure consisted of a reading habits questionnaire, containing six questions regarding participants' reading habits in everyday life, for each of which participants had to select one of five optional answers (adapted from Hartung, Burke, Hagoort, & Willems, 2016; e.g., *How often do you read fiction; How often do you read non-fiction; How many books do you read each year*). The indirect measure of reading habits was the Author Recognition Test (ART; Stanovich & West, 1989; Dutch adaptation reported in Koopman, 2015), consisting of 42 names (30 real authors and 12 foils), where participants had to indicate who they thought were genuine authors.

2.3.2.4 Personal characteristics (empathy, transportability)

To measure personal characteristics, such as transportability and empathy, participants filled out the Fantasy and Perspective Taking subscales of the Interpersonal Reactivity Index (IRI; Davis, 1980; Dutch translation adapted from De Corte et al., 2007) on a 7-point scale (e.g., *Becoming extremely involved in a good book or movie is somewhat rare for me; When I'm upset at someone, I usually try to "put myself in his shoes" for a while*). The Fantasy subscale measures the extent to which someone tends to get mentally involved in the stories they encounter, to the point at which they imagine themselves being part of the story (transportability). The Perspective Taking subscale measures the extent to which someone is able to take someone else's perspective in daily life.

2.4 Procedure

Participants first read the two stories in the MRI scanner, while their eye movements were being tracked. Stories were presented in counterbalanced order. Participants were instructed to read the stories the way they would also read for their own leisure. There was no additional task, and participants were able to proceed through the stories at their own pace. To proceed through the pages in the story, participants pressed a button with their right index finger when they finished reading a page. Both stories were divided into 30 pages. After each story, participants were allowed to take a short break from reading, to fill in the SWAS and appreciation questionnaire about the story they just read (while remaining inside the MRI scanner).

After reading the two stories, participants performed four localizer tasks (see Appendix B). Prior to this experiment, it was unknown whether individual differences in reading behavior would be detectable at the whole brain level. We therefore included the localizer tasks (now described in Appendix B) to have the opportunity to obtain functional regions of interest (ROIs). One important downside of these localizer tasks, however, was that they contain decontextualized stimuli, whereas our experiment and research question were specifically aimed at the neural processes that underlie naturalistic, contextualized reading (see Willems & Peelen, 2021 for a review about the differences in processing in response to contextualized versus decontextualized stimuli). Given that our whole brain analysis yielded functional ROIs that could be interrogated further, data obtained using the localizer tasks were not used. After the localizer tasks, participants left the MRI scanner, and completed the final questionnaires in a separate booth. First, participants answered the comprehension check questions about the two stories, after which the questions regarding reading habits and personal characteristics were asked.

No part of the study procedures was pre-registered prior to the research being conducted.

2.5 Stimulus Presentation

Stimuli were presented page by page on a projection screen (http://www.macada-innovision.nl) at the end of the bore, using a EIKI LC - XL100 beamer with a native resolution of 1024x768, with Presentation software (NBS, Berkeley, California). Participants could view the screen via a mirror (https://www.pgo-online.com/intl/katalog/cold-mirrors.html) mounted on the head coil. Pages consisted of maximally eight triple spaced lines. The distance between the mirror (110x100mm) and the projection screen (369x277mm) was 855 mm, and the distance between the mirror and the eye about 100mm (depending on how high a participant's head lies in the head coil).

2.6 Eye movement data acquisition and pre-processing

An MR compatible ceiling mounted Eyelink 1000 eye tracker (SR Research, Ottawa, Canada), with a sampling rate of 1000 Hz was used for eye movement data acquisition during scanning. The eye tracker

records infrared light reflected by the eyes, via a mirror attached to the head coil. The eye tracker was calibrated and calibration was validated before the presentation of each story.

Using SR Research's Eyelink Data Viewer, all fixations were checked before data analysis, and, if necessary, manually aligned. If this was impossible, because data were too noisy, data were excluded on a page-by-page basis. If too many pages had to be excluded within a participant, all data for this participant were excluded. This resulted in the exclusion of all data for three participants, the exclusion of one story for five participants (for one of these participants this was due to tracker malfunction rather than poor data quality), and in the exclusion of one to five pages in six participants (14 pages in total in these six participants). This amounts to a total of 14.33% of data loss based on eye tracking issues. After preprocessing, data for 37 participants were retained (full data for 27 participants, rejection of one story for five participants, rejection of a small portion of data for another five participants).

If entire story readings (or entire participants) needed to be removed due to poor eye tracking quality, the fMRI data for these story readings were also discarded (as we needed the eye tracking data to be able to analyze the fMRI data). In the cases where only one to five pages of eye tracking data needed to be removed, we did not discard the fMRI data for these participants. After preprocessing, we were able to use all fMRI data (two stories) for 32 participants, and fMRI data for one story for five participants. To be able to still analyze the fMRI data for the pages of which eye tracking data were discarded in the five participants for whom a small portion of the eye tracking data was rejected, we needed to impute the eye fixations for these data. To stay as close as possible to the participant's natural reading behavior, we modelled the onset and duration of the fixations, but imputed the mean value of the word characteristics we wanted to model as the weights of these fixations. This way the discarded part of the data would have no influence on the results of our analyses.

2.7 fMRI data acquisition and pre-processing

Data were collected at the Donders Centre for Cognitive Neuroimaging in Nijmegen, The Netherlands. fMRI data were acquired using a 3T MAGNETOM PrismaFit MR scanner (Siemens AG, Healthcare sector, Erlangen, Germany) with a 64-channel head-coil. Functional (TR = 1000ms, TE = 34ms, flip angle = 60˚, Field of View = 210mm, voxel size = 2.0x2.0x2.0mm, number of slices = 66, Multi-band acceleration factor = 6, multi-slice mode = interleaved, echo spacing = 0.62ms) and anatomical (Magnetization Prepared Rapid Acquisition Gradient Echo, voxel size = 1.0x1.0x1.0mm) images were acquired in one session lasting about 60 to 90 minutes, depending on the participants' reading speed.

Preprocessing was carried out using FEAT (version 6.00) in FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). The first ten or eleven volumes (ten or eleven seconds) were discarded (depending on the task programming) to allow for magnetic field saturation. Using FLIRT (Jenkinson,

Bannister, Brady, & Smith, 2002; Jenkinson & Smith, 2001), functional images were registered to the high resolution structural images (using Rigid-Body Transformation (6 DOF) and Boundary-Based Registration (Greve & Fischl, 2009)) after non-brain tissue was removed using BET (Smith, 2002). Motion correction was performed using MCFLIRT (Jenkinson et al., 2002), and values for the framewise displacement (average of rotation and translation parameter differences, using weighted scaling; Power, Barnes, Snyder, Schlaggar, & Petersen, 2012) as calculated using FSLMotionOutliers were saved as a confound EV for the first level analyses. High resolution structural images were registered to standard (MNI152 template, 2x2x2mm) space using FNIRT (Andersson, Jenkinson, & Smith, 2007a, 2007b) nonlinear registration (≥12 DOF). Spatial smoothing was performed using SUSAN noise reduction (Smith & Brady, 1997) with a 5mm FWHM Gaussian kernel. Grand-mean intensity normalisation of the entire 4D dataset was done by a single multiplicative factor. High pass temporal filtering was applied using Gaussian-weighted least-squares straight line fitting, with sigma=45.0s.

fMRI data preprocessing resulted in the additional exclusion of one story for one participant (1.25% data loss in addition to the data loss due to poor eye tracking quality). Note that there was a lot of overlap between the quality of the eye tracking and fMRI data: participants who move much during scanning, tend to have both poor eye tracking and poor fMRI data.

2.8 Data analysis

No part of the analyses was pre-registered prior to the research being conducted.

2.8.1 Eye tracking Data

The eye tracking data were analyzed in a similar way as the eye tracking data in Mak & Willems (2019) to make direct comparison of the eye tracking results possible. We analyzed how motor description, perceptual description and mental event descriptions related to gaze duration, while controlling for lexical frequency, word length and surprisal value as regressors of no interest, and allowing for random slopes and intercepts for the three types of descriptions over the interaction between subject and story to allow for individual variation between subjects and stories. Lexical frequency was derived from the SUBTLEX-NL database and consisted of the logarithm of the frequency with which a word appeared in the database (Keuleers, Brysbaert, & New, 2010). Word length was determined by counting the number of characters for each word. Surprisal value was derived from perplexity (i.e., an indication of the accuracy of a language model, in this case the perplexity is an indication of the accuracy with which a word is predicted by the previous words, cf. Lopopolo, Frank, van den Bosch, & Willems, 2017), calculated using a 3-gram model trained by SRILM (Stolcke, 2002) on 1 million sentences from the NLCOW2012 corpus (Schäfer & Bildhauer, 2012). Perplexity was equal to 10 to the power of negative surprisal. As in the analyses in Mak & Willems (2019), we used the values for the descriptions and word

characteristics of the previous word, as this allowed us to look in the spillover regions. We analyzed this with a Bayesian Multilevel Model using the package *brms* (Bürkner, 2017, 2018) and *Stan* (Stan Development Team, 2020) in *R* version 4.0.3 (R Core Team, 2021). The rationale for calculating a Bayesian multilevel model as opposed to a "classical" frequentist model, was that Bayesian models are more flexible and more capable of fitting complex models (e.g., Bürkner, 2018; Nalborczyk, Batailler, Lœvenbruck, Vilain, & Bürkner, 2019). Additionally, the analyses of the fMRI data were also done within a Bayesian framework (Beckmann, Jenkinson, & Smith, 2003; Woolrich, Behrens, Beckmann, Jenkinson, & Smith, 2004; Woolrich et al., 2009). Rather intuitively, Bayesian multilevel models calculate the range of the most probable values of each parameter, a 95% Credible Interval. If this Credible Interval does not cross zero for a given parameter, this indicates a 95% certainty that the true value of this parameter is distinguishable from zero.

In our model, we used weakly informative, normally-distributed priors with a mean of 0 and a standard deviation of 10 for all fixed effects. These priors are considered relatively conservative (McElreath, 2016). For the population-level intercept we used an informative, normally-distributed prior with a mean of 250 and a standard deviation of 50, since gaze durations are generally between 200 and 300 ms long on average. As variance can only be positive, weakly regularizing, half-cauchy priors with a mean of 0 and a standard deviation of 1 were used for the variance of the random effects as well as the overall variance (as suggested by Gelman, 2006; McElreath, 2016). The Gelman-Rubin diagnostic (*Rhat*) was 1.0 for all parameters (except for the intercept, for which it was 1.01), indicating that the model had converged.

2.8.2 fMRI Data

The fMRI-data were analyzed using a fixation-based analysis (comparable to an event-related analysis; see Richlan et al., 2014). In this analysis, the onset of a fixation was seen as the event onset, and the duration of the fixation as the event duration. These fixation events were then convolved with the HRF. From the eye-tracking data, we extracted the fixation (event) onsets and durations per word (which were determined automatically by SR Research's default parsing algorithm), to determine which word was looked at, at any given time during reading. Data analyses were performed in Feat (FMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). Time-series statistical analysis was carried out using FILM with local autocorrelation correction (Woolrich, Ripley, Brady, & Smith, 2001). For the first level analysis, we ran a GLM per participant (per run, one analysis for each story) where we modelled the onset and duration of each fixation, weighted by the scores for motor descriptions, perceptual descriptions, mental event descriptions, and the first principal

component[1] of lexical frequency, word length and surprisal value of the word that was fixated (to control for these word characteristics). This way we determined which brain areas respond specifically to either of the three types of descriptions in stories, while controlling for differences in word characteristics.

For statistical inference we contrasted each type of description with the other types of descriptions, to find the activation that was *specific* to that type of description (i.e., weighted contrasts [1 -.5 -.5] for motor > perceptual and mental event, perceptual > motor and mental event, mental event > motor and perceptual). Additionally, we contrasted each type of description with baseline (contrasts [1 0 0], [0 1 0] and [0 0 1], and visualized which areas were commonly activated by all three types of descriptions (as a conjunction analysis). The *z*-statistic images resulting from the contrasts were thresholded using clusters determined by $z > 3.1$ and a (corrected) cluster significance threshold of $p = .05$. The cluster threshold was determined based on the theory of Gaussian Random Fields (Worsley, 2001). This effectively controls the multiple comparisons problem introduced by the massive univariate approach taken at a family-wise error rate of $p < .05$ (Worsley, 2001).

As participants each read two stories, in two separate runs, we first aggregated the results for the two stories at the participant level using a standard weighted fixed effects model in FLAME (FMRIB Local Analysis of Mixed Effects). In this model, the variances from the first level analysis were used as the fixed effect error variances, and the random effects variance was forced to zero (Beckmann et al., 2003; Woolrich, 2008; Woolrich et al., 2004). The output from the fixed effects models per participant was used as input for the second level analysis. The second level analysis was performed using FLAME (FMRIB Local Analysis of Mixed Effects) stage 1 with automatic outlier detection, which estimates between-subject random effects using MCMC (Beckmann et al., 2003; Woolrich, 2008; Woolrich et al., 2004).

As a final, exploratory analysis, we investigated how individual differences in brain activation in response to the different types of descriptions are related to the experience of narrative reading as well as individual difference measures. We did this to find out whether brain activation due to simulation-eliciting content occurs equally or differently across individuals and to find out whether any individual differences could be explained by reading experiences or personal characteristics. The analysis was done by first extracting the percent signal change (per participant and per story, from the first level analyses) in five or six regions of interest for each of the three types of descriptions. We selected

[1] To avoid multicollinearity issues due to the high correlation between lexical frequency, word length and surprisal value, we entered the first principal component of these variables into the model instead of entering each of the variables seperately into the model. For this principal component analysis (PCA), the Kaiser-Meyer-Olkin measure (KMO) was .71 (all KMO values for individual items > .65), indicating good sampling adequacy for this analysis. Bartlett's test of sphericity showed sufficient correlation between items, $\chi 2\ (3) = 597.40$, $p < .001$. The scree-plot in combination with the eigenvalues found in an initial analysis (Kaiser's method) and the model fit (fit based upon off diagonal values) confirmed that it was appropriate to summarize the three word characteristics into one principal component. This principal component explained 84% of the variance.

regions of interest that were (1) significantly activated by one or all of the three types of descriptions on the group level, that were (2) part of large clusters of activation in response to the descriptions, and that were (3) good candidates for finding individual differences in simulation based on previous literature (e.g., Chow et al., 2015; Grill-Spector & Weiner, 2014; Igelström & Graziano, 2017; Kurby & Zacks, 2013; Moody & Gennari, 2010; Nijhof & Willems, 2015). We derived the regions of interest from the results of our group analysis: we extracted the by-participant by-story percent signal change from areas that were found to be commonly activated by these descriptions. We then built models to predict percent signal change in each area, by scores on the Story World Absorption Scale, the appreciation questionnaire, the Fantasy and Perspective taking subscales of the Interpersonal Reactivity Index, the Author Recognition Test, and the questions about reading habits (see heading "Questionnaires" below for more information on these questionnaires). We built separate models for Story World Absorption and for appreciation, to make sure that any conceptual overlap between the two would not skew our results. We analyzed this with Bayesian Multilevel Models using the package *brms* (Bürkner, 2017, 2018) and *Stan* (Stan Development Team, 2020) in *R* version 4.0.3 (R Core Team, 2021). In our models, we used weakly informative, normally-distributed priors with a mean of 0 and a standard deviation of 1 for all fixed effects. For the population-level intercept we used a weakly informative, normally-distributed prior with a mean of 0 and a standard deviation of 10. These priors are considered relatively conservative (McElreath, 2016). As variance can only be positive, weakly regularizing, half-cauchy priors with a mean of 0 and a standard deviation of 1 were used for the variance of the random effects as well as the overall variance (as suggested by Gelman, 2006; McElreath, 2016). In all models, the Gelman-Rubin diagnostic (*Rhat*) was 1.0 for all parameters, indicating that the models had converged.

3. Folder 1: bids defaced MRI data

This folder contains the defaced MRI data, per participant subfolder. It contains the defaced anatomical data (.nii.gz, .json) in the subfolder “anat”, and the functional data (.nii.gz, .json) per scanner run in the subfolder “func”. The file “Key sub-alias.xlsx” in the DAC explains which run belongs to which task for each of the participants. The folder also contains relevant metadata.

4. Folder 2: DataViewerSessions

This folder contains information on the **eye tracking data analysis.** This folder contains 41 subfolders for the participants (40 experimental participants and 1 pilot participant – BlitP01), in which the preprocessed eye tracking data (.evs) can be found for each of the stories. Additionally, it contains IAS MDALB and IAS S_S, containing the interest area sets of story 1/MDALB and story 2/S&S/Signs for the purpose of the eye tracking analysis. Finally, it contains files with the output of the interest area reports from DataViewer, for each of the stories, and a file detailing which pages/data sets in the eye-tracking data were adjusted or rejected, how, and why.

Note that the participant aliases in the eye-tracking data (BLitP**) are different from the participant numbers in the MRI data (sub-0**). The file “Key sub-alias.xlsx” in the RDC explains which participant numbers belong to which participant alias.

5. Folder 3: EventFiles

This folder contains the event files for all scanner runs. The event files detail the succession of all events in the experimental tasks (e.g., when were stimuli presented, for how long).

6. Folder 4: eyetracker texts

This folder contains the texts of the two stories, split up over 30 screens each, as they were presented in the scanner. These are the images (.jpg) that were presented using the Presentation scripts (see folder 3), and they can be imported into the raw eyetracking data files to see where on each page fixations landed. Files containing “MDALB” in their filename, belong to story 1 “De mensen die alles lieten bezorgen”, pages 1-30. Files containing “S&S” in their filename, belong to story 2 “Signalen en Symbolen”, pages 1-30.

7. Folder 5: raw_behavioral

This folder contains the raw behavioral data per participant, for each task. There are eyetracking files (.edf), containing the raw eye tracking data (except for task BP (button press), since no eyetracking was registered in this task). Note that within the eye tracking data files, the participants are named differently than in the subfolder names. The file “Key sub-alias.xlsx” in the DAC explains which Alias belongs to which subject. Apart from the edf-files, there are logfiles (.log), detailing the exact timing and succession of the events in the scanner run. Tasks are BP (button press localizer), EyeLoc (eye movement localizer), MDALB (story 1), S&S (story 2), SimLoc (simulation localizer), and WL (wordlist).

8. References

Andersson, J. L. R., Jenkinson, M., & Smith, S. M. (2007a). *Non-linear optimisation. FMRIB technical report TR07JA1*. Oxford.

Andersson, J. L. R., Jenkinson, M., & Smith, S. M. (2007b). *Non-linear registration, aka Spatial normalisation. FMRIB technical report TR07JA2*. Oxford.

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, *20*, 1052–1063. https://doi.org/10.1016/S1053-8119(03)00435-X

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80–84. https://doi.org/10.1016/j.actpsy.2014.04.010

Bürkner, P.-C. (2017). brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395. https://doi.org/10.32614/RJ-2018-017

Chow, H. M., Mar, R. A., Xu, Y., Liu, S., Wagage, S., & Braun, A. R. (2015). Personal experience with narrated events modulates functional connectivity within visual and motor systems during story comprehension. *Human Brain Mapping*, *36*, 1494–1505. https://doi.org/10.1002/hbm.22718

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85.

De Corte, K., Buysse, A., Verhofstadt, L. L., Roeyers, H., Ponnet, K., & Davis, M. H. (2007). Measuring empathic tendencies: Reliability and validity of the Dutch version of the Interpersonal Reactivity Index. *Psychologica Belgica*, *47*(4), 235–260. https://doi.org/10.5334/pb-47-4-235

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548. https://doi.org/10.1038/nrn3747

Hartung, F., Burke, M., Hagoort, P., & Willems, R. M. (2016). Taking perspective: Personal pronouns affect experiential aspects of literary reading. *PLoS ONE*, *11*(5), 1–18. https://doi.org/10.1371/journal.pone.0154732

Igelström, K. M., & Graziano, M. S. A. (2017). The inferior parietal lobule and temporoparietal junction: A network perspective. *Neuropsychologia*, *105*, 70–83. https://doi.org/10.1016/j.neuropsychologia.2017.01.001

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1016/S1053-8119(02)91132-8

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. https://doi.org/10.1016/S1361-8415(01)00036-6

Jobe, T. (2009). Power Analysis for mixed-effect models in R. Retrieved from https://toddjobe.blogspot.com/2009/09/power-analysis-for-mixed-effect-models.html

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Knoop, C. A., Wagner, V., Jacobsen, T., & Menninghaus, W. (2016). Mapping the aesthetic space of literature "from below." *Poetics*, *56*, 35–49. https://doi.org/10.1016/j.poetic.2016.02.001

Koopman, E. M. (Emy). (2015). Empathic reactions after reading: The role of genre, personal factors and affective responses. *Poetics*, *50*, 62–79. https://doi.org/10.1016/j.poetic.2015.02.008

Kuijpers, M. M., Hakemulder, F., Tan, E. S., & Doicaru, M. M. (2014). Exploring absorbing reading experiences: Developing and validating a self-report scale to measure story world absorption. *Scientific Study of Literature*, *4*(1), 89–122. https://doi.org/10.1075/ssol.4.1.05kui

Kurby, C. A., & Zacks, J. M. (2013). The activation of modality-specific representations during discourse processing. *Brain and Language*, *126*, 338–349. https://doi.org/10.1016/j.bandl.2013.07.003

Lopopolo, A., Frank, S. L., van den Bosch, A., & Willems, R. M. (2017). Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLOS ONE*, *12*(5), e0177794. https://doi.org/10.1371/journal.pone.0177794

Mak, M., Faber, M., & Willems, R. M. (n.d.). Different routes to liking: How readers arrive at narrative evaluations.

Mak, M., & Willems, R. M. (2019). Mental simulation during literary reading: Individual differences revealed with eye-tracking. *Language, Cognition and Neuroscience*, *34*(4), 511–535. https://doi.org/10.1080/23273798.2018.1552007

McElreath, R. (2016). *Statistical Rethinking*. Boca Raton, Florida: Chapman and Hall/CRC. https://doi.org/10.1201/9781315372495

Moody, C. L., & Gennari, S. P. (2010). Effects of implied physical effort in sensory-motor and pre-frontal cortex during language comprehension. *NeuroImage*, *49*(1), 782–793. https://doi.org/10.1016/j.neuroimage.2009.07.065

Nabokov, V. (2003). Signalen en Symbolen. In *Een Russische schoonheid 1*. Amsterdam: De Bezige Bij.

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Nijhof, A. D., & Willems, R. M. (2015). Simulating fiction: Individual differences in literature comprehension revealed with fMRI. *PLoS ONE*, *10*(2), 1–17. https://doi.org/10.1371/journal.pone.0116492

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

R Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Richlan, F., Gagl, B., Hawelka, S., Braun, M., Schurz, M., Kronbichler, M., & Hutzler, F. (2014). Fixation-related fMRI analysis in the domain of reading research: Using self-paced eye movements as markers for hemodynamic brain responses during visual letter string processing. *Cerebral Cortex*, *24*, 2647–2656. https://doi.org/10.1093/cercor/bht117

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M. Ugur Dogan, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012* (pp. 486–493).

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062

Smith, S. M., & Brady, J. M. (1997). SUSAN—A New Approach to Low Level Image Processing. *International Journal of Computer Vision*, *23*, 45–78. https://doi.org/https://doi.org/10.1023/A:1007963824710

Speed, L. J., & Brysbaert, M. (2022). Dutch sensory modality norms. *Behavior Research Methods*, *54*(3), 1306–1318. https://doi.org/10.3758/s13428-021-01656-9

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*(4), 402–433. https://doi.org/10.2307/747605

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing, ICSLP 2002*.

Team, S. D. (2020). RStan: the R interface to Stan. Retrieved from http://mc-stan.org

Van Essen, R. (2014). De mensen die alles lieten bezorgen. In *Hier wonen ook mensen* (pp. 113–123). Utrecht: Atlas Contact.

Willems, R. M., & Peelen, M. V. (2021). How context changes the neural basis of perception and language. *IScience*, *24*, 102392. https://doi.org/10.1016/j.isci.2021.102392

Woolrich, M. W. (2008). Robust group analysis using outlier inference. *NeuroImage*, *41*(2), 286–301. https://doi.org/10.1016/j.neuroimage.2008.02.042

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, *21*(4), 1732–1747. https://doi.org/10.1016/j.neuroimage.2003.12.023

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., … Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, *45*, S173–S186. https://doi.org/10.1016/j.neuroimage.2008.10.055

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data. *NeuroImage*, *14*(6), 1370–1386. https://doi.org/10.1006/nimg.2001.0931

Worsley, K. J. (2001). Statistical analysis of activation images. In P. Jezzard, P. M. Matthews, & S. M. Smith (Eds.), *Functional Magnetic Resonance Imaging* (pp. 251–270). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780192630711.003.0014