



Analyzing the Realization of Recipients

Utku Turk¹

¹ Department of Linguistics, Boğaziçi University

Author Note

Utku Turk, M.A. student at the Department of Linguistics, Boğaziçi University.

Correspondence concerning this article should be addressed to Utku Turk, John Freely Building, Department of Linguistics, South Campus, Boğaziçi University. E-mail:

utku.turk@boun.edu.tr




Abstract

Up until today, the introspection procedure has been extensively used **without ever being questioning**. This procedure allows linguists to gather observational data which has empirical status to some extent. However, presenting a set of sentences and expecting native judgments on these sentences can be quite misleading. **The sentences prepared by the researcher may transform participants into subjects, forcing them to serve the researcher's intuition**. Because corpora have become widespread and are easy to analyze thanks to recent developments in computer science, many researchers have started to use daily linguistic data without setting limitations. In this paper, a dataset **explicating** the details of the use of the dative structure in the Switchboard corpus and the Treebank Wall Street Journal collection has been used to **illustrate** the effects of the definiteness and the accessibility of both the recipient and the theme on the realization of the dative structure in English sentences.


Keywords: R, Ditransitive predicates, NP Realization, Alienization Effect

Analyzing the Realization of Recipients



The Problem



Big data and the advanced use of statistical tools give rise to an one of the intriguing **question**: How do people form ditransitive predicates and which factors determine the internal structure of the verb phrase? Traditionally, these kinds of grammatical structures have been analyzed from a more theoretical perspective and have focused on native judgments, thus making it impossible to differentiate the results from the researcher's own intuition. Moreover, many studies have indeliberately shown that judgments are extremely problematic, and most of the time these judgments come from an **awfully** restricted group of people, a bell-jar around the researcher. Both of these problematic situations are exemplified **to a great extent** by studies on dative alternation (Bresnan, Cueni, Nikita, & Baayen, 2007), which is the main focus of this paper.




In this paper, the main question I asked and set out to find a **sensible** explanation for is as follows: What determines the phrasal structure of ditransitive predicates in English? Throughout the paper and the data analysis, two distinct characteristics of both the recipient and the theme are used to identify when and why native speakers of English choose to use prepositional dative structures. The independent variables in this analysis are the definiteness of recipient, the definiteness of theme, the accessibility of theme, and the accessibility of recipient. With these variables, I aim to explain the **effect of Alienization** on dative structures.




The Dataset


The dataset used in this paper is from the R package languageR named dative from the study by Bresnan, Cueni, Nikita, and Baayen (2007). From this data set, I have selected 5 columns to focus on and have described the properties of the $N = 3263$ observation in the Switchboard corpus and the Treebank Wall Street Journal collection. The selected columns from this dataset form the basis of this paper's analysis.




Definitions

 In this section, I will provide definitions for important keywords I use throughout the paper. First, explaining the main data observed is of utmost importance. According to **Martin** Haspelmath (2013), ditransitive verbs are verbs with two arguments in addition to the subject: a “recipient” or “addressee” argument, and a “theme” argument. While the recipient is a special kind of goal where the action is directed towards and which is associated with verbs expressing a change in ownership, the theme is the element that undergoes the action but does not change its state. (Dowty, 1991)

The definiteness of a phrase is determined with other elements that precede it in the dataset. Certain determiners such as *a/an*, *many*, *some*, and *either* mark an NP as indefinite whereas others, including *the*, *this*, *every*, and *both* mark an NP as definite. (Huddleston & Pullum, 2002) In the dataset, accessibility columns consist of three unique types of information: given, new, and accessible. These columns identify the context accessibility of the recipient and the theme. *New* accessibility implies that the element uttered is newly introduced to the discourse, *given* means that it was already uttered in the interaction, and *accessible* means even though it is not explicitly introduced to the discourse, it is available in the discourse.

The term *Alienization* I use refers to the conceptual distance that is created upon uttering the elements of the sentence. In line with the Relevance Principle, which puts forth that the closer an affix is to the verb root, the more relevant it is to the meaning (Greenberg & Social Science Research Council. Linguistics and Psychology Committee., 1966), the  concept of alienization implies that phrases that are immediately preceded by the verb are more relevant in terms of discourse regarding ditransitive structures in English.

What to Expect

Before starting to fit a model, I will demonstrate the dataset  **in the form of lineplots** in order to show the relevant relationships and to offer a better understanding of what to

expect and what not to expect.

Definiteness

As can be seen in *Figure 1*, the percentage of NP realization is affected significantly by both the definiteness of theme and definiteness of Recipient. While theme definiteness decreases the chance of NP realization of the recipient, definiteness of the recipient increases the percentage. Moreover, it is readily observable that the differences due to definiteness are coequally portrayed, thus nullifying the possibility of a significant interaction between these two predictors.

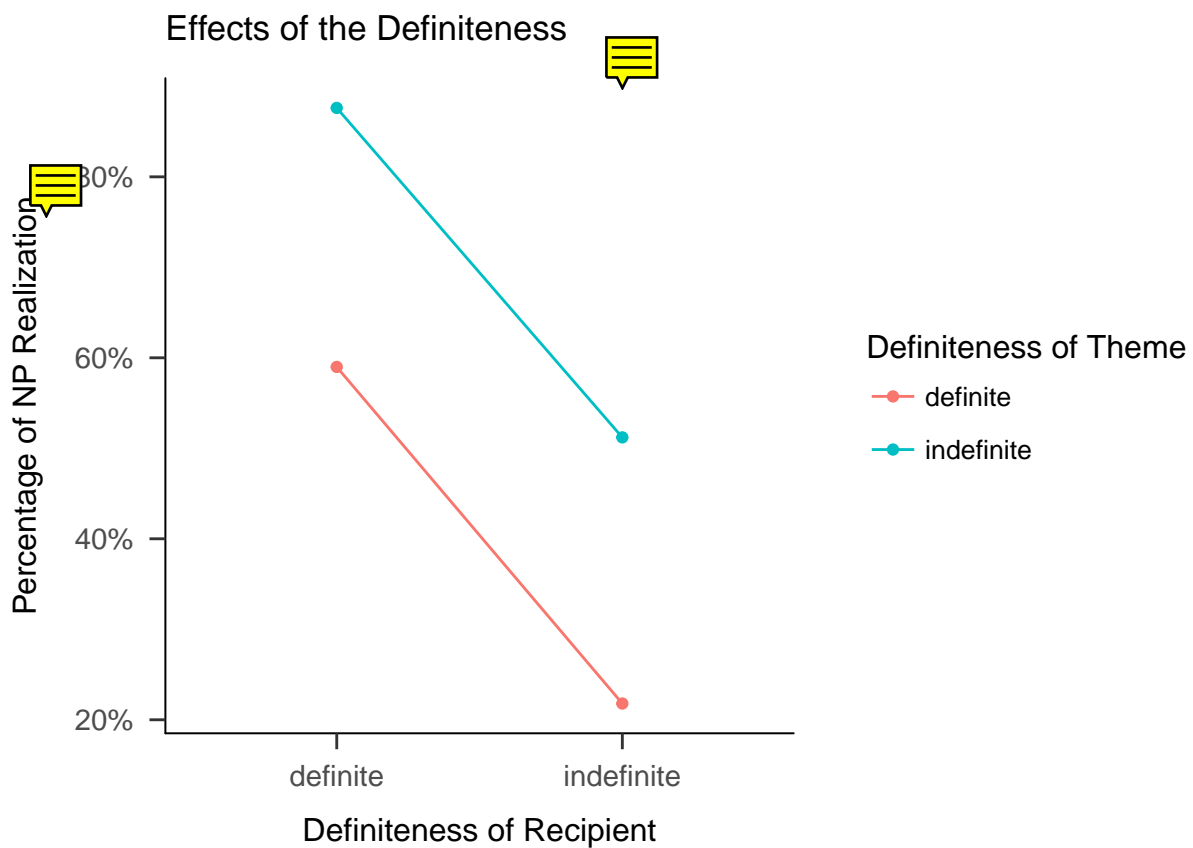


Figure 1

Accessibility

As for the *Effects of Accessibility*, while the accessible and new theme do not really differ in terms of the percentage of NP realization, the given theme definitely and rather substantially decreases the NP realization tendency. The most plausible scenario for NP Realization is when the theme is new, and the recipient is given with the percentage of $M = .951$.

Even though the percentage of NP realization is more pronounced in the *accessible* recipient, the somewhat stable *given* theme line tells us that it is a property of theme, instead of an interaction.

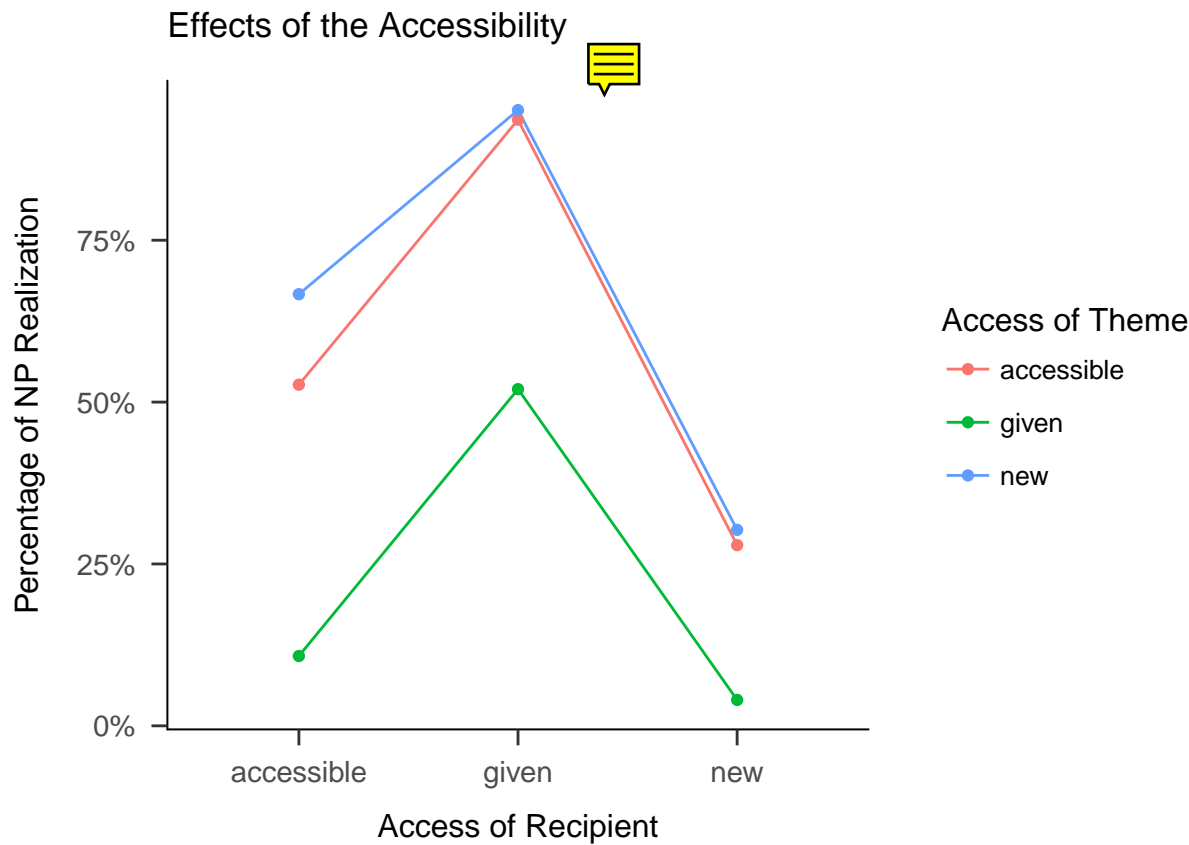


Figure 2

Recipient



The effects related to the characteristics of the recipient are shown in *Figure 3*. It follows my first hypothesis which implied that there would be a tendency toward an NP realization if the recipient was more relevant in the discourse. Looking at *Figure 3*, one can easily argue that when the recipient is *given* in the discourse and is definite, English speakers put the recipient in the primary position, making it a noun phrase rather than a prepositional phrase. Interaction between the Access and the Definiteness of the recipient is not non-existent, yet I do not think it is significant to the extent that it will create a bigger effect with other dependent variables.

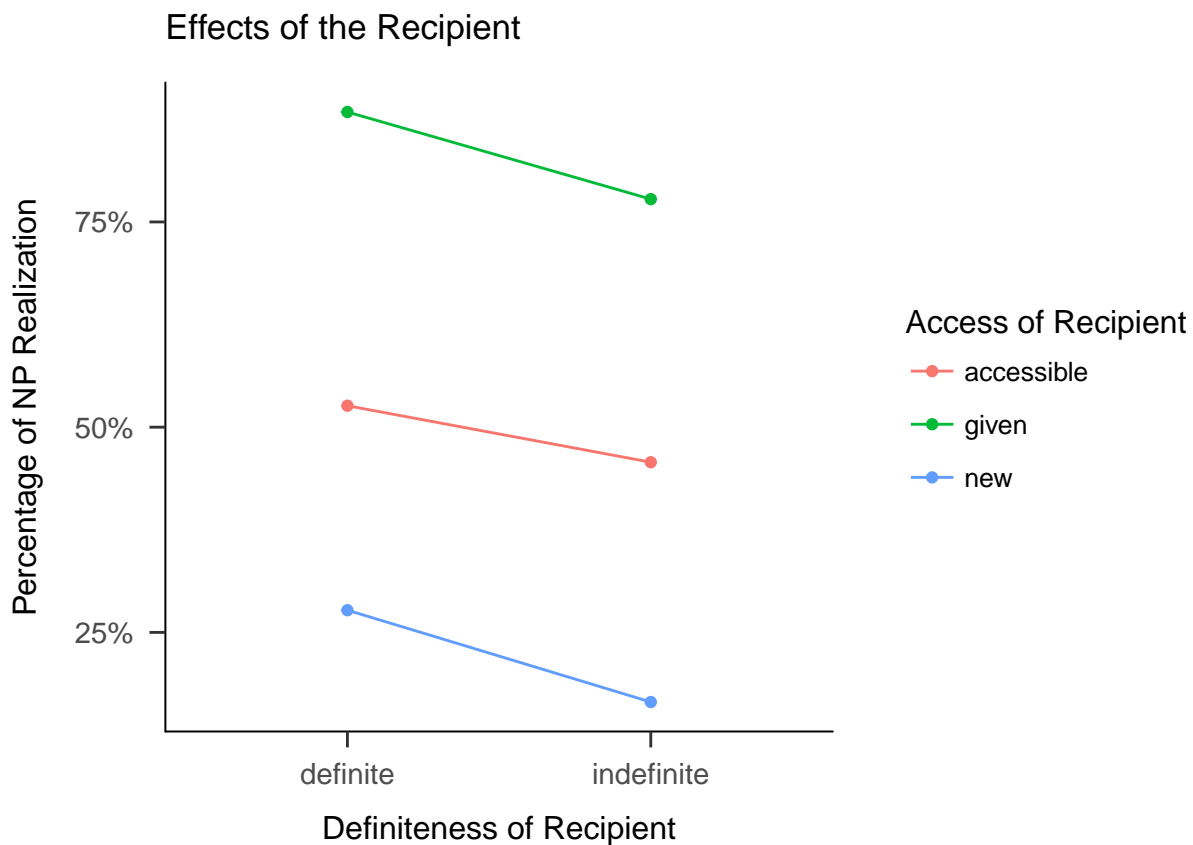


Figure 3

Theme



The percentage of NP realizations in the data as a function of the access and the definiteness of theme is shown in *Figure 4*. Among the four plots, this one clearly stands out. When the theme is definite, the NP Realization Hierarchy follows as such: *Accessible* > *New* > *Given*. Yet, this order is reversed completely when the theme is indefinite. What we would normally expect is that when the theme is definite, it should have a higher tendency to use an NP in a new discourse, yet this is not the case here.

However, looking at the numbers in the accessible discourse, we can easy infer that there is a somewhat fixed percentage. There must be an interaction that tilts the percentages as such.

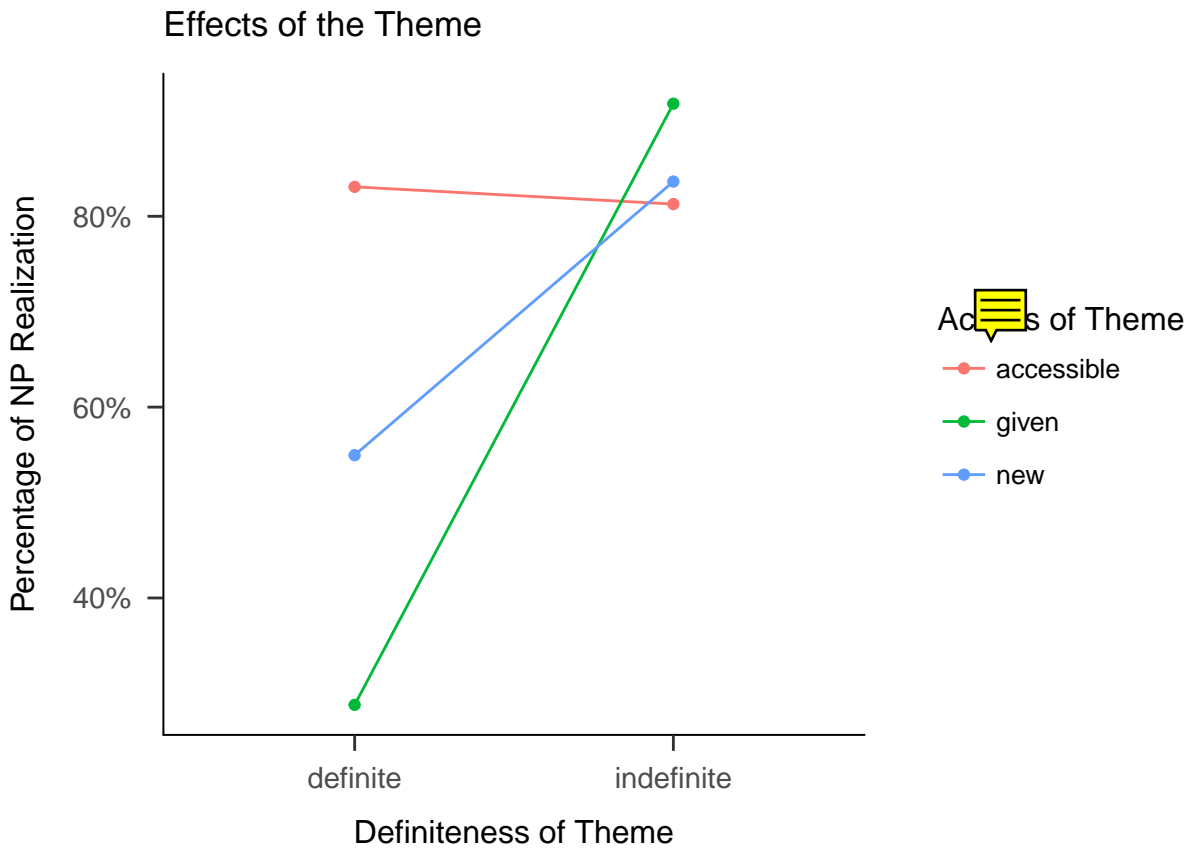


Figure 4



Methods

In the dataset, I included only 5 columns regarding the data, which are discussed in the paper. The purpose of the paper is to identify the effects of those variables; the rest is not included. Moreover, while treatment contrasts are used for the plots, sum contrasts are utilized in the model.

Apart from the relative information provided in the section titles **Dataset**, I used R (Version 3.4.2; R Core Team, 2017) and the R-packages *bindrcpp* (Version 0.2; Müller, 2017), *brms* (Version 2.0.1; Bürkner, 2017), *citr* (Version 0.2.0; Aust, 2016), *coda* (Version 0.19.1; Plummer, Best, Cowles, & Vines, 2006), *dplyr* (Version 0.7.4; Wickham, Francois, Henry, & Müller, 2017), *ggplot2* (Version 2.2.1; Wickham, 2009), *languageR* (Version 1.4.1; Baayen, 2013), *lazerhawk* (Version 0.1.9; Clark, n.d.), *magrittr* (Version 1.5; Bache & Wickham, 2014), *pander* (Version 0.6.1; Daróczi & Tsegelskyi, 2017), *papaja* (Version 0.1.0.9655; Aust & Barth, 2017), and *Rcpp* (Eddelbuettel & Balamuta, 2017; Version 0.12.14; Eddelbuettel & François, 2011) for the analyses, figures, and the tables provided in the paper.

Procedure and Data Analysis

Having explained the dataset, problem, and what to expect from the dataset, we can advance to our model. First, the display of the model used in the paper can be found below.

$$Realization_i \sim Bernoulli(\mu_i)$$

$$logit(\mu_i) = \alpha + \gamma_k \times DoR_k + \beta_{AoR_k} \times AoR + \gamma_i \times DoT_i + \beta_{AoT} \times AoT_i$$

$$\gamma_i = \beta_{DoT} + \beta_{AoTDoT}AoT_i$$

$$\gamma_k = \beta_{DoR} + \beta_{AoRDoR}AoR_k$$

In the model specified above, we used two linear models, which leads to a logistic regression. The first line defines the likelihood function I used; it is a Bernoulli distribution with logit link, which is specified in the second line. The likelihood function consists of additive

definition integrated with another additive definition γ_i which is a placeholder for the linear function that defines the slope between *Definiteness of Theme* and *Access of Theme*. *DoR*, *AoR*, *DoT*, *AoT* stands for the definiteness of recipient, the accessibility of recipient, the definiteness of theme, and the definiteness of theme, respectively. Also, γ_k is another placeholder for a similar function this time between the relevant characteristics of recipient. Each β stands for the coefficient relevant to the indeendent variable. This model is run through (Bürkner, 2017), which sets improper flat priors by default. These priors are not changed. Further analyses may be focused on identifying the relevant priors and using them.

All models, interpretations, and functions are run through and interpreted via Bayes Theorem. The primary underlying motivation behind the utilization of Bayes Theorem is the fact that it provides me with interpretable and reproducible answers without any fee except the computational power of my processor. Even though the dataset's sample size is rather large, reducing Bayes Theorem's importance to some degree, describing and updating the probabilities of my hypothesis given the evidence carry utmost importance for this paper. Since the Bayesian Analysis complies with my likelihood function, the additive evidence, and allows me to use a computationally-rich MCMC model, I ran such and analysis and interpreted the output using a Bayesian approach.

Results

The summary of our model is specified below in *Table 1*. Rows in the table shows names of covariates, their estimates, their 95% credible intervals, and whether their credible interval contain 0 or not. Looking at the *Table 1*, our model suggest stringkingly different things we thought in the first place. Either the model is misspecified and needed to be revised or there is no significant effect of definiteness and the accessibility as our credible internal not only contains 0 but almost in a fifty-fifty situation.

Even though the accessible recipient is not notable, meaning contains 0 in its credible interval, it is only -.01 to .34, thus, making it 34 times more likely to have a positive slope

Table 1



Descriptive statistics of NP Realizations



Covariate	Estimate	Est.Error	l-95% CI	u-95% CI	Notable
Model Intercept	-0.01	0.13	-0.23	0.20	
Definite Recipient	-0.39	0.09	-0.53	-0.24	*
Accessible Recipient	0.17	0.11	0.00	0.34	
Given Recipient	-1.73	0.15	-1.99	-1.48	*
Definite Theme	0.79	0.11	0.62	0.97	*
Accessible Theme	-0.42	0.12	-0.62	-0.22	*
Given Theme	0.72	0.20	0.39	1.03	*
Definite & Accessible Recipient	0.25	0.11	0.07	0.41	*
Definite & Given Recipient	-0.27	0.15	-0.51	-0.01	*
Definite & Accessible Theme	-0.97	0.12	-1.18	-0.78	*
Definite & Given Theme	1.13	0.20	0.82	1.46	*

than the negative one.

Discussion


Looking at the raw *Table 2*, we can see the the *Highest Posterior Density Intervals* of the covariates we have. Our model demonstrates that definiteness of the recipient, recipient given in the discourse, accessible theme, and the interaction between the definite theme and given theme decreases tendency of using noun phrases for the recipient in English given the data for sure, meaning their credible intervals do not contain zero. Apart from those, the interaction between definite recipient and given recipient also decreases the the tendency of percentage of NP realizations.

At this point, I think writing a new model or having more narrow analysis would do

better, yet for the purpose of this first draft, I will try to explain as much as I can. ##

Marginal Effects What our intercept means? Our intercept means the tendency of having NP Recipients when the definiteness values are unweighted and the accessibility values are



new, which is kind of interesting. It actually says something profoundly meaningful. *New* theme and recipient cannot win over the other, making the NP realizations of recipients almost arbitrary ($M = -0.01$). 

Thus, all the marginal estimates shown in the *Table 1* is the effects of those on the *New Theme and Recipient Rivalry*. Therefore, those are not the effects of what is stated on the NP Recipient Realization, rather they are effects of the NP Recipient Realization in the framework of *New Theme and Recipient Rivalry*.

With this in mind, what I expected from the model is that the maximum NP Realization of the recipient is when the recipient is definite, theme is indefinite, recipient is given, and the theme is given. However, upon computing the mean of posterior samples of what I thought would give me the maximum NP realization, I realize that I am at the ground zero again ($M = -3.189$)

At this point, I am again unable to discuss anything before to have any insight on the data I have been working on for 2 weeks until I have redesigned my model and thought about it.

Table 2

Highest Posterior Density Intervals

	lower	upper
b_Intercept	-0.25	0.26
b_DefinOfRec1	-0.57	-0.21
b_AccessOfRec1	-0.04	0.38
b_AccessOfRec2	-2.02	-1.43
b_DefinOfTheme1	0.59	1.01
b_AccessOfTheme1	-0.67	-0.19
b_AccessOfTheme2	0.35	1.10
b_DefinOfRec1:AccessOfRec1	0.03	0.45
b_DefinOfRec1:AccessOfRec2	-0.56	0.03
b_DefinOfTheme1:AccessOfTheme1	-1.20	-0.73
b_DefinOfTheme1:AccessOfTheme2	0.77	1.54
lp___	-1,175.48	-1,166.83

References

- Aust, F. (2016). *Citr: 'RStudio' add-in to insert markdown citations*. Retrieved from <https://CRAN.R-project.org/package=citr>
- Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from <https://github.com/crsh/papaja>
- Baayen, R. H. (2013). *LanguageR: Data sets and functions with “analyzing linguistic data: A practical introduction to statistics”*. Retrieved from <https://CRAN.R-project.org/package=languageR>
- Bache, S. M., & Wickham, H. (2014). *Magrittr: A forward-pipe operator for r*. Retrieved

- from <https://CRAN.R-project.org/package=magrittr>
- Bresnan, J., Cueni, A., Nikita, T., & Baayen, R. H. (2007). Predicting the dative alternation. *Cognitive Foundations of Interpretation*, 69–94.
- Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. doi:[10.18637/jss.v080.i01](https://doi.org/10.18637/jss.v080.i01)
- Clark, M. (n.d.). *Lazerhawk: Miscellaneous functions mostly inspired by synthwave*. Retrieved from m-clark.github.io
- Daróczi, G., & Tsegelskyi, R. (2017). *Pander: An r 'pandoc' writer*. Retrieved from <https://CRAN.R-project.org/package=pander>
- Dowty, D. R. (1991). Montague's General Theory of Languages and Linguistic Theories of Syntax and Semantics. In (pp. 1–36). doi:[10.1007/978-94-009-9473-7_1](https://doi.org/10.1007/978-94-009-9473-7_1)
- Eddelbuettel, D., & Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, 5, e3188v1. doi:[10.7287/peerj.preprints.3188v1](https://doi.org/10.7287/peerj.preprints.3188v1)
- Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8), 1–18. doi:[10.18637/jss.v040.i08](https://doi.org/10.18637/jss.v040.i08)
- Greenberg, J. H., & Social Science Research Council. Linguistics and Psychology Committee. (1966). *Universals of language; report of a conference held at Dobbs Ferry, New York, April 13-15, 1961. Edited by Joseph H. Greenberg.* (p. 337). M.I.T. Press.
- Huddleston, R. D., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language* (p. 1842). Cambridge University Press. Retrieved from <http://www.cambridge.org/tr/academic/subjects/languages-linguistics/grammar-and-syntax/cambridge-grammar-english-language{\#}XkWUSrQepi4REuRv.97>
- Martin Haspelmath. (2013). Ditransitive Constructions: The Verb 'Give'. Retrieved from <http://wals.info/chapter/105>
- Müller, K. (2017). *Bindrcpp: An 'rcpp' interface to active bindings*. Retrieved from

<https://CRAN.R-project.org/package=bindrcpp>

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1), 7–11. Retrieved from

<https://journal.r-project.org/archive/>

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from

<https://www.R-project.org/>

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Retrieved from <http://ggplot2.org>

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *Dplyr: A grammar of data manipulation*. Retrieved from <https://CRAN.R-project.org/package=dplyr>