Modelling dependency completion in sentence comprehension as a Bayesian hierarchical mixture process: A case study involving Chinese relative clauses

Shravan Vasishth (vasishth@uni-potsdam.de)
Department of Linguistics, University of Potsdam, Germany, and
Centre de Recherche en Mathématiques de la Décision, CNRS, UMR 7534, Université Paris-Dauphine, PSL Research University, Paris, France.

Nicolas Chopin (nicolas.chopin@ensae.fr)
École Nationale de la Statistique et de l'administration économique, Malakoff, France.

Robin Ryder (ryder@ceremade.dauphine.fr)
Centre de Recherche en Mathématiques de la Décision, CNRS, UMR 7534, Université Paris-Dauphine, PSL Research University, Paris, France.

Bruno Nicenboim (bruno.nicenboim@uni-potsdam.de)
Department of Linguistics, University of Potsdam, Germany.

Abstract

In sentence comprehension, it is widely assumed (Gibson 2000, Lewis & Vasishth, 2005) that the distance between linguistic co-dependents affects the latency of dependency resolution: the longer the distance, the longer the retrieval time (the distance-based account). An alternative theory of dependency resolution difficulty is the direct-access model (McElree et al., 2003); this model assumes that retrieval times are a mixture of two distributions: one distribution represents successful retrieval and the other represents an initial failure to retrieve the correct dependent, followed by a reanalysis that leads to successful retrieval. The time needed for a successful retrieval is independent of the dependency distance (cf. the distance-based account), but reanalyses cost extra time, and the proportion of failures increases with increasing dependency distance. We implemented a series of increasingly complex hierarchical Bayesian models to compare the distance-based account and the direct-access model; the latter was implemented as a hierarchical finite mixture model with heterogeneous variances for the two mixture distributions. We evaluated the models using two published data-sets on Chinese relative clauses which have been used to argue in favour of the distance account, but this account has found little support in subsequent work (e.g., Jäger et al., 2015). The hierarchical finite mixture model, i.e., an implementation of direct-access, is shown to provide a superior account of the data than the distance account.

Keywords: Bayesian hierarchical Finite Mixture Models; Psycholinguistics; Sentence Comprehension; Chinese Relative Clauses; Direct-Access Model

Introduction

In sentence comprehension research, dependency completion is assumed by many theories to be a key event. For example, in a sentence like *The man slept*, in order to understand who did what, the subject noun is retrieval at the verb. One well-known generalization is that dependency distance partly determines comprehension difficulty as measured by reading times or question-response accuracy. For example, the Dependency Locality Theory discussed in Gibson and Wu (2013), and the cue-based retrieval account of Lewis and Vasishth (2005) both assume that the longer the distance between two co-dependents such as a subject and a verb, the greater the retrieval difficulty at the moment of dependency completion. We will call this the *distance account*.

As an example, consider the self-paced reading study in Gibson and Wu (2013). The dependent variable here was the reading time at the head noun (*official*) in Chinese subject and object relative clauses. As shown in (1), in Chinese subject relatives, the distance is larger between the head noun and the gap it is coindexed with (the coindexing is marked with the subscript i), compared to object relatives.[1] For simplicity, we operationalize distance here as the number of words intervening between the gap inside the relative clause and the head noun.

(1) a. Subject relative
[GAP$_i$ yaoqing fuhao de] guanyuan$_i$
GAP invite tycoon DE official
xinhuaibugui
have bad intentions
'The official who invited the tycoon has bad intentions.

b. Object relative
[fuhao yaoqing GAP$_i$ de] guanyuan$_i$
tycoon invite GAP DE official
xinhuaibugui
have bad intentions
'The official who the tycoon invited has bad intentions.

In the Gibson and Wu study, reading times were recorded using self-paced reading in the two conditions (subject relative, coded $-1/2$, and object relative, coded $+1/2$), with 37 subjects and 15 items, presented in a standard Latin square

[1] The dependency could be equally well be between the relative clause verb and the head noun; nothing hinges on assuming a gap-head noun dependency.

design. The experiment originally had 16 items, but one item was removed in the published analysis due to a mistake in the item.

The distance account's predictions can be evaluated by fitting the hierarchical linear model shown in (1). Assume that (i) i indexes participants, $i = 1,\ldots,I$ and j indexes items, $j = 1,\ldots,J$; (ii) y_{ij} is the reading time in milliseconds for the i-th participant reading the j-th item; and (iii) the predictor X is sum-coded ($\pm 1/2$). Then, the data y_{ij} (reading times in milliseconds) are defined to be generated by the following model:

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + b_i + c_j + \varepsilon_{ij} \quad (1)$$

where $b_i \sim Normal(0,\sigma_b^2)$, $c_j \sim Normal(0,\sigma_c^2)$ and $\varepsilon_{ij} \sim Normal(0,\sigma_e^2)$. The terms b_i and c_j are called varying intercepts for participants and items respectively; they represent by-subject and by-item adjustments to the fixed-effect intercept β_0. Their variances, σ_b^2 and σ_c^2 represent between-participant (respectively item) variance.[2] This model is effectively a statement about how the data are assumed to be generated. If the distance account is correct, we would expect to find evidence that the slope β_1 is statistically significantly different from zero; specifically, reading times for subject relatives are expected to be longer than those for object relatives. As shown in Table 1, this prediction is borne out. Subject relatives are estimated to be read 120 ms slower than object relatives, apparently consistent with the predictions of the distance-based account.

	Estimate	Std. Error	t value
(Intercept)	548.43	51.56	10.64
$X_{\text{subj-rel:}-1/2}$	-120.39	48.01	-2.51*

Table 1: A linear mixed model using raw reading times in milliseconds as dependent variable, corresponding to the reported results in Gibson and Wu 2013.

In summary, the theoretical interpretation of this finding, originally presented in Gibson and Wu (2013), is that in Chinese, subject relatives are harder to process than object relatives because the gap inside the relative clause is more distant from the head noun in subject vs. object relatives. This makes it more difficult to complete the gap-head noun dependency in subject relatives. This distance-based explanation of processing difficulty is plausible given the considerable independent evidence from languages such as English, German, Hindi, Persian and Russian that dependency distance can affect reading time (see review in Safavi, Husain, and Vasishth (2016)).

[2]This so-called crossed participants and items varying intercepts linear mixed model can be made more complex by adding varying slopes for the factor X by participant and by item, but for ease of exposition we do not consider these more complex models in the present paper. Such complex models would anyway be overparameterized given the amount of data available.

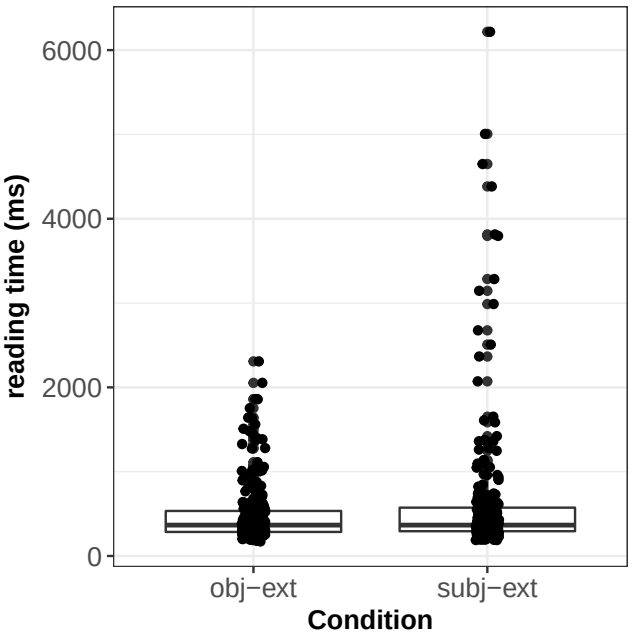

Figure 1: Boxplots showing the distribution of reading times by condition of the Gibson and Wu (2013) data.

However, not only has recent work on Chinese relatives (Jäger, Chen, Li, Lin, & Vasishth, 2015) cast doubt on dependency distance account, even in the data considered by Gibson and Wu, the distributions of the reading times for the two conditions show an interesting asymmetry that cannot be straightforwardly explained by the distance account. The reading times in subject relatives are much more spread out than in object relatives. This is shown in Figure 1. Although this spread was ignored in the original analysis, a standard response to heterogeneous variances (heteroscedasticity) is to delete "outliers" based on some criterion; a common criterion is to delete all data lying beyond $\pm 2.5SD$ in each condition. This procedure assumes that the data points identified as extreme are irrelevant to the question being investigated. An alternative approach is to not delete data but to downweight the extreme values by applying a variance stabilizing transform (Box & Cox, 1964). Taking a log-transform of the reading time data, or a reciprocal transform, can reduce the heterogeneity in variance; see Vasishth, Chen, Li, and Guo (2013) for analyses of the Gibson and Wu data using a transformation.

However, it is possible that the heteroscedasticity in subject and object relatives in the Gibson and Wu data reflects a systematic difference in the underlying generative processes of reading times in the two relative clause types. We investigate this question by modelling the extreme values.

Using the probabilistic programming language Stan (Stan Development Team, 2016), we show that a hierarchical mixture model provides a better fit to the data (in terms of predictive accuracy) than several simpler hierarchical models.

The mixture model that we present below can be seen as a model of extreme values. As Nicenboim and Vasishth (2016a) pointed out, the underlying generative process is consistent with an implementation of the direct-access model of McElree, Foraker, and Dyer (2003). We therefore suggest that, at least for the Chinese relative clause data, the direct-access model may be a better way to characterize the dependency resolution process than the distance-based account.

Reading times as a mixture distribution: The direct-access model of McElree et al. (2003)

A finite mixture model assumes that the outcome (here, reading time in milliseconds, $y_i, i = 1,\ldots,N$) is drawn from one of several distributions.[3] Each distribution's identity is controlled by a categorical mixing distribution. For example, assume that we have K distributions with location parameter (the mean) $\mu_k \in \mathbb{R}$ and scales (standard deviation) $\sigma_k \in (0,\infty)$. If they are mixed in proportions $\lambda =< \lambda_1,\ldots,\lambda_K >$, where $\lambda_k \geq 0$ and $\sum_{k=1}^{K} \lambda_k = 1$, for each outcome y_i there is a latent variable $z_i \in \{1,\ldots,K\}$ with a categorical distribution parameterized by $\lambda : z_i \sim \texttt{Categorical}(\lambda)$. The dependent variable y_i (reading time in milliseconds), is then distributed as a Log-Normal distribution (for further justification, see Nicenboim and Vasishth (2016b):

$$y_i \sim LogNormal(\mu_{z_i}, \sigma^2_{z_i}) \tag{2}$$

As mentioned above, the direct-access model can be seen as assuming a mixture distribution where successful retrieval can be modelled as a Log-Normal distribution: $y \sim LogNormal(\mu, \sigma^2_e)$. Retrieval failure and subsequent reanalysis can be modelled as another Log-Normal distribution with a different location parameter and possibly also a different scale parameter: $y \sim LogNormal(\mu+\delta, \sigma^2_{e'})$.

Thus, one can consider the direct-access model as a mixture model with retrieval time as generated from one of two distributions, where the proportion of trials in which a retrieval failure occurs (the mixing proportion) is p:

$$\begin{aligned} y \sim & p \cdot LogNormal(\mu+\delta, \sigma^2_{e'}) \\ & +(1-p) \cdot LogNormal(\mu, \sigma^2_e) \end{aligned} \tag{3}$$

In order to understand whether the Chinese relative clause data are best described as being generated by a mixture process, we implemented a series of increasingly complex models and compared the relative fit of these models. All models were hierarchical, with varying intercepts for participant and for item.

Model comparison using Pareto-smoothed importance sampling Leave-One-Out (PSIS-LOO) cross-validation

Model comparison can be carried out using different methods; here, we use an approximation of the leave-one-out cross-validation (LOO) approach, as discussed in Vehtari, Gelman, and Gabry (2016b). In essence, LOO compares the expected predictive performance of alternative models by subsetting the data into a training set (for estimating parameters) by excluding one observation. The difference between the predicted and observed held-out value can then be used to quantify model quality by successively holding out each observation. The sum of the expected log pointwise predictive density, $\widehat{elpd}$, can be used as a measure of predictive accuracy, and the difference between the $\widehat{elpd}$'s of competing models can be computed, including the standard deviation of the sampling distribution of the difference in $\widehat{elpd}$. When comparing a model M0 with another model M1, if M1 has a higher $\widehat{elpd}$, then it has a better predictive performance compared to M0. The quantity $\widehat{elpd}$ is a Bayesian alternative to the Akaike Information Criterion (Akaike, 1974). Vehtari and colleagues developed an efficient computation of LOO using Pareto-smoothed importance sampling (PSIS-LOO), and this is what we use here. Details of PSIS-LOO are omitted here due to space constraints; see Vehtari et al. (2016b) for more. We used the LOO package, version 1.0.0 (Vehtari, Gelman, & Gabry, 2016a) to compute the $\widehat{elpd}$ for each model.

Definitions of the hierarchical mixture models

The different hierarchical models evaluated are shown in Table 2. The dependent variable is reading time in milliseconds, and the underlying generating distribution for the reading times is a Log-Normal; the varying intercepts for subjects and items are assumed to be normally distributed, as is standard in linear mixed models. Priors are defined for the model parameters as follows. All standard deviations are constrained to be greater than 0 and have priors $\texttt{Cauchy}(0, 2.5)$; probabilities have priors $\texttt{Beta}(1,1)$; and all coefficients (β parameters) have priors $\texttt{Cauchy}(0, 2.5)$.

In the mixture models, we will call the distribution that corresponds to the successful retrieval the *success distribution*, and the one corresponding to the retrieval failure followed by a reanalysis the *failure distribution*.

We fit six models, described below and shown more formally in Table 2. Other models can be fit too, but are omitted here due to space restrictions.

- M0: A standard linear mixed model (no mixture). This corresponds to a test of the distance-based account.
- M1: A mixture only in subject relatives (homogeneous variance). This model assumes that failures happen only in subject relatives (SRs).
- M2: A mixture only in subject relatives (heterogeneous variance). Here, we assume, as in M1, that failures happen only in SRs, but the variance of the failure distribution is assumed to be different than the variance of the success distribution.
- M3: This model assumes that there is no difference in retrieval time in ORs vs SRs, but only in the probability of

[3] We follow the presentation of mixture models in Stan Development Team (2016).

Model	Definition	Variables
M0	$y_{ij} \sim LogNormal(\beta_0 + \beta_1 X_{ij} + b_i + c_j, \sigma_e^2)$	$b_i \sim Normal(0, \sigma_b^2)$, $c_j \sim Normal(0, \sigma_c^2)$; and σ_e^2 is the variance of the distribution corresponding to a correct retrieval.
M1	SRs: $y_{ij} \sim p \cdot LogNormal(\beta_{SR} + b_i + c_j + \delta, \sigma_e^2)$ $+(1-p) \cdot LogNormal(\beta_{SR} + b_i + c_j, \sigma_e^2)$ ORs: $y_{ij} \sim LogNormal(\beta_{OR} + b_i + c_j, \sigma_e^2)$	p is the probability of failure. β_{SR} and β_{OR} are the means for subject and object relatives. Thus, `diff` = $\beta_{SR} - \beta_{OR}$. δ is the additional cost in the failure distribution.
M2	SRs: $y_{ij} \sim p \cdot LogNormal(\beta_{SR} + b_i + c_j + \delta, \sigma_{e'}^2)$ $+(1-p) \cdot LogNormal(\beta_{SR} + b_i + c_j, \sigma_e^2)$ ORs: $y_{ij} \sim LogNormal(\beta_{OR} + b_i + c_j, \sigma_e^2)$	$\sigma_{e'}^2$ is the variance of the failure distribution.
M3	SRs: $y_{ij} \sim p_{SR} \cdot LogNormal(\beta + b_i + c_j + \delta, \sigma_e^2)$ $+(1-p_{SR}) \cdot LogNormal(\beta + b_i + c_j, \sigma_e^2)$ ORs: $y_{ij} \sim p_{OR} \cdot LogNormal(\beta + b_i + c_j + \delta, \sigma_e^2)$ $+(1-p_{OR}) \cdot LogNormal(\beta + b_i + c_j, \sigma_e^2)$	p_{SR}, p_{OR} are failure probabilities in SRs (ORs). `diffprob` is the difference in failure probability in SR vs OR (`diffprob`=$p_{SR} - p_{OR}$). β is the common mean reading time for subject and object relatives.
M4	SRs: $y_{ij} \sim p_{SR} \cdot LogNormal(\beta + b_i + c_j + \delta, \sigma_{e'}^2)$ + $(1-p_{SR}) \cdot LogNormal(\beta + b_i + c_j, \sigma_e^2)$ ORs: $y_{ij} \sim p_{OR} \cdot LogNormal(\beta + b_i + c_j + \delta, \sigma_{e'}^2)$ $+(1-p_{OR}) \cdot LogNormal(\beta + b_i + c_j, \sigma_e^2)$	$\sigma_{e'}$ and σ_e are the SDs of the two distributions. For other variables, see above.
M5:	See above	As model M4, except that separate β parameters are assumed for SRs and ORs.

Table 2: The model definitions. The best model among these for the Gibson and Wu 2013 data is M4.

successful retrieval. The variances of the success and failure distributions are assumed to be identical (homogeneous variances).

- M4: This model also assumes that there is no difference in retrieval time in ORs vs SRs, but only in the probability of successful retrieval. Unlike M3, the variances of the success and failure distributions are assumed to be different (heterogeneous variances).
- M5: This model assumes that retrieval time in SRs and ORs is different, and that the variances of the two distributions are different (heterogeneous variance). Thus, M5 is like M4, but with the additional assumption that distance may affect dependency completion time.

The data

The evaluation of these models was carried out using two separate data-sets. The first was the original study from (Gibson & Wu, 2013) that was discussed in the introduction. The second study was a replication of the Gibson and Wu study that was published in Vasishth et al. (2013). This second study served the purpose of validating whether independent evidence can be found for the mixture model selected using the original Gibson and Wu data.

Results

The original Gibson and Wu study As shown in Table 3, a comparison of the models M0-M5 using LOO show that (a) the model M1, which assumes a mixture only in subject relatives with homogeneous variances in the mixture distributions, is better than the standard linear mixed model M0; (b) the model M2, which assumes a mixture only in subject relatives but with heterogeneous variances in the mixture distributions, is no better than the simpler model M1; (c) M3, which assumes mixture distributions in both subject and object relatives but homogeneous variances, outperforms M2; (d) M4, which assumes mixture distributions in both subject and object relatives but heterogeneous variances, outperforms M3; (e) M4 has a numerically better fit (lower $\widehat{elpd}$) than the more complex model M5, which assumes different means for SRs and ORs.

Among the models considered, the best model—the one with the lowest $\widehat{elpd}$—is therefore the heterogeneous variance mixture model M4. Posterior predictive checks also confirm that M4 generates simulated data that is consistent with the observed data, but the simpler model M0 fails to generate the spread observed in the subject relatives in the Gibson and Wu 2013 data (details omitted due to space limitations).

Table 4 shows, for the two mixture models M3 and M4, the posterior estimates of the parameters, along with 95% credible intervals (these demarcate the range of plausible values of the parameter with probability 95%). We display only models M3 and M4 because these two models approximate the direct-access model of McElree et al. (2003).

From Table 4 we can see that in both M3 and M4, in subject

relatives the failure distribution occurs with a higher probability than in object relatives (0.25 vs 0.21). In model M4, the mean difference in the probability of the failure distribution is 4%, with a 95% credible interval [-5,13]%. The posterior probability of this difference being greater than zero is 82%.

Furthermore, since M4 is the selected model, the failure distribution is better characterized as having a larger variance ($\hat{\sigma}_{e'} = 0.64$ in M4) than the success distribution ($\hat{\sigma}_e = 0.22$ in M4).

comp	elpd_diff	se
M0 vs M1	56.31	15.49
M1 vs M2	1.00	1.96
M2 vs M3	60.87	17.81
M3 vs M4	29.99	9.42
M5 vs M4	-29.79	4.32

Table 3: Comparison of the different models for the Gibson and Wu 2013 data using PSIS-LOO.

models	parameter	mean	lower	upper
m3	beta	5.90	5.79	6.02
	delta	1.35	1.26	1.44
	diffprob	0.04	-0.02	0.10
	prob_sr	0.14	0.10	0.19
	prob_or	0.10	0.07	0.14
	sigma_e	0.30	0.28	0.32
	sigma_u	0.25	0.19	0.33
	sigma_w	0.13	0.08	0.20
m4	beta	5.85	5.75	5.95
	delta	0.93	0.72	1.14
	diffprob	0.04	-0.05	0.13
	prob_sr	0.25	0.18	0.34
	prob_or	0.21	0.14	0.29
	sigmap_e	0.64	0.53	0.75
	sigma_e	0.22	0.20	0.25
	sigma_u	0.24	0.18	0.31
	sigma_w	0.09	0.05	0.16

Table 4: Parameter estimates from the homogeneous variance mixture model (M3) and heterogeneous variance mixture model (M4), with 95 percent credible intervals.

The replication of the Gibson and Wu study In this data-set, there were 40 participants and the same 15 items as in Gibson and Wu's data were used. Figure 2 shows the distribution of the data by condition; there seems to a similar skew as in the original study, although the spread is not as dramatic as in the original study.

Table 5 shows that M4 is again the best model. Table 6 shows that we obtain parameter estimates in this replication data-set for model M4 that are similar to those from the original Gibson and Wu data. In particular, the mean difference in the probability of the failure distribution for subject vs. object relatives is 7%, with a 95% credible interval [-1,16]%.

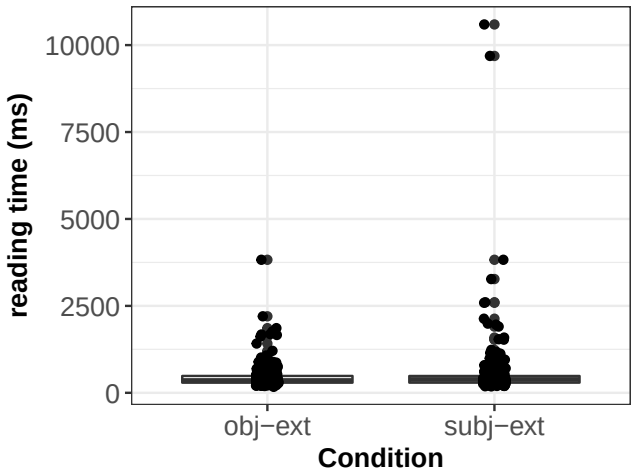

Figure 2: Boxplots showing the distribution of reading times by condition of the replication of the Gibson and Wu data.

comp	elpd_diff	se
M0 vs M1	67.75	22.00
M1 vs M2	17.56	11.31
M2 vs M3	13.84	28.67
M3 vs M4	55.91	18.24
M5 vs M4	-63.81	13.32

Table 5: Comparison of the different models in the replication data using PSIS-LOO.

	Original data			Replication data		
param.	mean	lower	upper	mean	lower	upper
beta	5.85	5.76	5.95	5.86	5.78	5.95
delta	0.93	0.74	1.14	0.74	0.55	0.96
diffprob	0.04	-0.05	0.13	0.07	-0.01	0.16
prob_sr	0.25	0.18	0.33	0.23	0.15	0.33
prob_or	0.21	0.14	0.29	0.16	0.09	0.25
sigmap_e	0.64	0.54	0.75	0.69	0.59	0.81
sigma_e	0.22	0.20	0.25	0.21	0.18	0.23
sigma_u	0.24	0.18	0.31	0.22	0.17	0.29
sigma_w	0.09	0.05	0.16	0.07	0.03	0.12

Table 6: The posterior distributions of the parameters from the mixture model M4 using the original Gibson and Wu 2013 data and the replication data from Vasishth et al. 2013.

The posterior probability of this difference being greater than zero is 95%. The replication data thus suggest a systematic difference in retrieval failures occurs in subject versus object relatives.

Discussion

The model comparison and parameter estimates presented above suggest that, at least as far as the Chinese relative clause data are concerned, a better way to characterize the dependency completion process is in terms of the direct-access

model and not the distance account implied by Gibson and Wu (2013) and Lewis and Vasishth (2005). Specifically, there is suggestive evidence in the Gibson and Wu (2013) data that a higher proportion of retrieval failures occurred in subject relatives compared to the object relatives. In other words, increased dependency distance may have the effect that it increases the proportion of retrieval failures (followed by reanalysis).

There is one potential objection to the conclusion above. It would be important to obtain independent evidence as to which dependency was eventually created in each trial. This could be achieved by asking participants multiple-choice questions to find out which dependency they built in each trial. Although such data is not available for the present study, in other work (on number interference) (Nicenboim, Engelmann, Suckow, & Vasishth, 2016) did collect this information. There, too, we found good evidence for the direct-access model (Nicenboim & Vasishth, 2016a). In future work on Chinese relatives, it would be helpful to carry out a similar study to determine which dependency was completed in each trial. In the present work, the modelling at least shows how the extreme values in subject relatives can be accounted for by assuming a two-mixture process.

Conclusion

The mixture models suggest that, in the specific case of Chinese relative clauses, increased processing difficulty in subject relatives is not due to dependency distance leading to longer reading times, as suggested by Gibson and Wu (2013). Rather, a more plausible explanation for these data is in terms of the direct-access model of McElree et al. (2003). Under this view, retrieval times are not affected by the distance between co-dependents, but a higher proportion of retrieval failures occur in subject relatives compared to object relatives. This leads to a mixture distribution in both subject and object relatives, but the proportion of the failure distribution is higher in subject relatives.

Acknowledgments

We are very grateful to Ted Gibson for generously providing the raw data and the experimental items from Gibson and Wu (2013). Thanks also go to Lena Jäger for many insightful comments. For partial support of this research, we thank the Volkswagen Foundation through grant 89 953.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

Gibson, E., & Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. *Language and Cognitive Processes*, *28*(1-2), 125–155.

Jäger, L. A., Chen, Z., Li, Q., Lin, C.-J. C., & Vasishth, S. (2015). The subject-relative advantage in Chinese: Evidence for expectation-based processing. *Journal of Memory and Language*, *79–80*, 97–120. doi: 10.1016/j.jml.2014.10.005

Lewis, R. L., & Vasishth, S. (2005, May). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*, 67–91.

Nicenboim, B., Engelmann, F., Suckow, K., & Vasishth, S. (2016). *Number interference in German: Evidence for cue-based retrieval.* (submitted to Cognitive Science)

Nicenboim, B., & Vasishth, S. (2016a). *Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling.* Retrieved from `https://arxiv.org/abs/1612.04174` (arXiv eprint)

Nicenboim, B., & Vasishth, S. (2016b). Statistical methods for linguistic research: Foundational ideas – Part II. *Language and Linguistics Compass*, *10*, 591-613.

Safavi, M. S., Husain, S., & Vasishth, S. (2016). Dependency resolution difficulty increases with distance in Persian separable complex predicates: Implications for expectation and memory-based accounts. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.00403

Stan Development Team. (2016). Stan modeling language users guide and reference manual, version 2.12 [Computer software manual]. Retrieved from `http://mc-stan.org/`

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013, 10). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE*, *8*(10), 1–14.

Vehtari, A., Gelman, A., & Gabry, J. (2016a). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models [Computer software manual]. Retrieved from `https://github.com/stan-dev/loo` (R package version 1.0.0)

Vehtari, A., Gelman, A., & Gabry, J. (2016b). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*.