Comparing Theories of Speaker Choice Using a Model of Classifier Production in Mandarin Chinese

Comparing Theories of Speaker Choice Using a Model of Classifier Production in Mandarin Chinese

Meilin Zhan
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
meilinz@mit.edu

Roger Levy
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
rplevy@mit.edu

Abstract

Speakers often have more than one way to express the same meaning. What general principles govern speaker choice in the face of optionality when near semantically invariant alternation exists? Studies have shown that optional reduction in language is sensitive to contextual predictability, such that the more predictable a linguistic unit is, the more likely it is to get reduced. Yet it is unclear to what extent these cases of speaker choice are driven by audience design versus toward facilitating production.

Here we argue that for a different optionality phenomenon, namely classifier choice in Mandarin Chinese, Uniform Information Density and at least one plausible variant of availability-based production make opposite predictions regarding the relationship between the predictability of the upcoming material and speaker choices. In a corpus analysis of Mandarin Chinese, we show that the distribution of speaker choices supports the availability-based production account, and not Uniform Information Density.

1 Introduction

The expressivity of natural language often gives speakers multiple ways to convey the same meaning. Meanwhile, linguistic communication takes place in the face of environmental and cognitive constraints. For instance, language users have limited memory and cognitive resources, the environment is noisy, and so forth. What general principles govern speaker choice in the face of alternations that are (nearly) semantically invariant? To the extent that we are able to provide a general answer to this question it will advance our fundamental knowledge of human language production.

Studies have shown that alternations are very often sensitive to contextual predictability. For well-studied cases of optional REDUCTION in language, the following trend is widespread: the more predictable a linguistic unit is, the more likely it is to get reduced. Predictable words are phonetically reduced (Jurafsky et al., 2001; Bell et al., 2009; Seyfarth, 2014) and have shorter lexical forms (Piantadosi et al., 2011), and optional function words are more likely to be omitted when the phrase they introduce is predictable (Jaeger, 2010). Yet it is unclear to what extent speakers' choices when faced with an alternation are made due to audience design or to facilitate production. For example, the above pattern of predictability sensitivity in optional reduction phenomena is predicted by both the Uniform Information Density (UID) hypothesis (Levy and Jaeger, 2007), a theory which that the speaker aims to convey information at a relatively constant rate and which can be motivated via considerations of optimality from the comprehender's perspective (e.g., Smith and Levy, 2013), and by the speaker-centric availability-based production hypothesis (Bock, 1987; Ferreira and Dell, 2000), which hypothesizes that the dominant factor in determining speaker choice is that the speaker uses whatever material is readily available when it comes time to convey a particular part of a planned message.

Here we argue that for a different optionality phenomenon, namely classifier choice in Mandarin Chinese, UID and availability-based production make opposite predictions regarding the relationship between the predictability of upcoming material and speaker choice. In a corpus analysis of Mandarin Chinese, we show that the distribution of speaker choices supports the availability-based production account, and not UID.

Proceedings of NAACL-HLT 2018, pages 1997–2005
New Orleans, Louisiana, June 1 - 6, 2018.

2 Uniform Information Density and Availability-based Production

In Sections 2 and 3, we explain why the UID and availability-based production accounts make the same predictions in many cases, but can be potentially disentangled using Chinese classifier choice. Here we exemplify predictions of these two accounts in the case of optional function word omission.

For optional function word omission such as *that*-omission ((1) and (2)), predictability effects have been argued to be consistent with both the speaker-oriented account of AVAILABILITY-BASED PRODUCTION (Bock, 1987; Ferreira and Dell, 2000) and the potentially audience-oriented account of UNIFORM INFORMATION DENSITY (Levy and Jaeger, 2007). On both accounts, but for different reasons, the less predictable the clause introduced by the functional word, the more likely the speaker will be to produce the function word *that*.

(1) The student (that) you tutored graduated.

(2) The woman thought (that) we were crazy.

The UID hypothesis claims that within boundaries defined by grammar, when multiple options are available to encode a message, speakers prefer the variant that distributes information density most uniformly, thus lowering the chance of information loss or miscommunication (Levy and Jaeger, 2007; Jaeger, 2010). In (1), if the function word *that* is omitted, the first word of the relative clause *you* serves two purposes: signaling the onset of the relative clause, and conveying part of the contents of the relative clause itself. These both contribute to the information content of the first relative clause-internal word. If one or both is high-surprisal, then the first relative clause-internal word might be a peak in information density, as illustrated in Figure 1 (top left). If instead the function word *that* is produced, *that* signals the onset of the relative clause, and *you* only communicates part of the content of the relative clause itself. This could help eliminate any sharp peak in information density, as illustrated in Figure 1 (bottom left). Thus, if the speaker's goal is to transfer information as smoothly as possible, the less predictable the upcoming clause, the more inclined the speaker would be to produce the function word *that*.

On the availability-based production hypothesis, speaker choice is governed by the relationship by the relative time-courses of (i) when a part of a message needs to be expressed within an utterance, and (ii) when the linguistic material to encode that part of the message becomes available for production. If material that specifically encodes a part of the message is available when it comes time to convey that part of the message, it will be used—that is the PRINCIPLE OF IMMEDIATE MENTION of Ferreira and Dell (2000). If, on the other hand, that material is not yet available, then other available material consistent with the grammatical context produced thus far and that does not cut off the speaker's future path to conveying the desired content will be used. In (1), assuming the function word *that* is always available when the speaker plans to produce a relative clause, the speaker will produce *that* when the upcoming relative clause or the first part of its contents are not yet available. If phrase structures and phrase contents take longer to become available when they are lower-predictability—an assumption consistent with the literatures on picture naming (Oldfield and Wingfield, 1965) and word naming (Balota and Chumbley, 1985)—then the less predictable the relative clause, the lower the probability that its first word, w_1, will be available when the time comes to begin the relative clause, as illustrated in Figure 2 (left). Under these circumstances, the speaker would choose to produce other available material, namely function word *that*. If, in contrast, the upcoming relative clause is predictable, then w_1 will be more likely to be available, and the speaker would be more likely to omit the function word *that* and immediately proceed with w_1.

While these two accounts differ at many levels, they make the same prediction for function word omission in syntactic reduction such as (1) and (2). It is difficult to disentangle these accounts empirically.[1] Below we will show that for a different optionality phenomenon, namely classifier choice in Mandarin, these accounts may make different predictions.

[1] Prior work (Jaeger, 2010) acknowledged this entanglement of the predictions of these accounts, and attempted to tease the accounts apart via joint modeling using logistic regression. The present study builds on these efforts by exploring a case involving a starker disentanglement of the accounts' predictions.

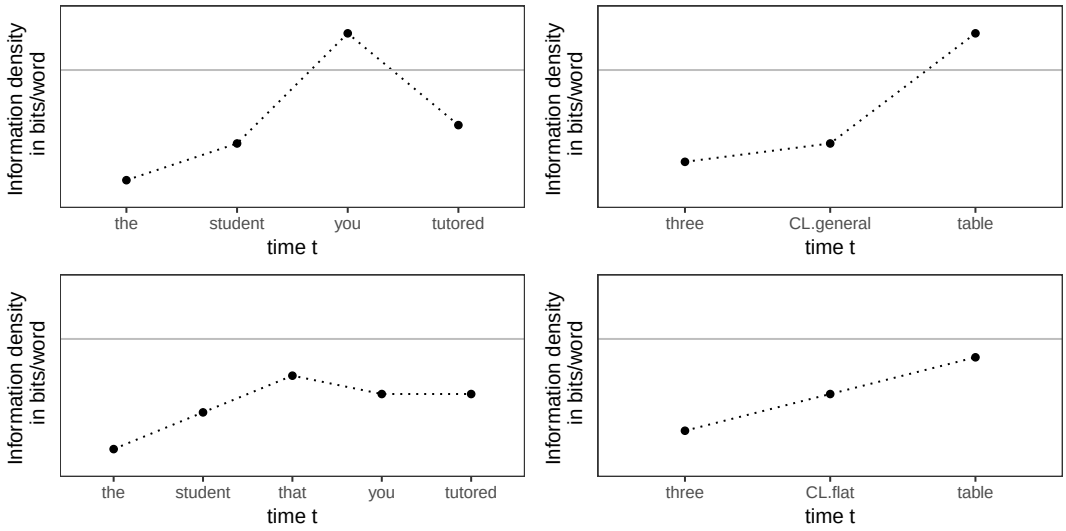

Figure 1: Schematic illustrations of Uniform Information Density in the context of relative clause (left) and classifier choice (right). The grey lines indicate a hypothetical channel capacity.

3 Classifiers in Mandarin Chinese

Languages in the world can be broadly grouped into classifier languages and non-classifier languages. In non-classifier languages, such as English and other Indo-European languages, a numeral modifies a noun directly: e.g., *three tables, two projects*. In Mandarin Chinese and other classifier languages, a numeral classifier is obligatory when a noun is to be preceded with a numeral (and often obligatory with demonstratives): e.g., *san zhang zhuozi* "three CL.flat table", *liang xiang gongcheng* "two CL.item project". Although it has been hypothesized that numeral classifiers play a functional role analogous to that of the singular–plural distinction in other languages (Greenberg, 1972), it is not clear whether there is a meaningful correlation between the presence of numeral classifiers and plurality among the languages of the world (Dryer and Haspelmath, 2013).

In Mandarin, classifiers, together with their associated numeral or demonstrative, precede the head noun of a noun phrase. There are about 100 individual numeral classifiers (Ma, 2015). While different nouns are compatible with different SPECIFIC classifiers, there is a GENERAL classifier *ge*(个) that can be used with most nouns. In some cases, the alternating options between using a general or a specific classifier with the same noun are almost semantically invariant. Table 1 shows examples of classifier options in fragments of naturally occuring texts.

Yet these options have different effects on the information densities of the following nouns. A specific classifier is more likely to reduce the information density of the upcoming noun than a general classifier because a specific classifier constrains the space of possible upcoming nouns more tightly (Klein et al., 2012). Consider the following pair of classifier examples (3) and (4).

(3) 我 买了 三 张 桌子
wo mai-le san zhang zhuozi
I bought three CL.flat table ("I bought three tables")

(4) 我 买了 三 个 桌子
wo mai-le san ge zhuozi
I bought three CL.general table ("I bought three tables")

As shown in Figure 1 (top right), while a general classifier has some information (e.g., signaling there will be a noun), it has relatively low information density—it is the most frequent and generally the highest-probability classifier in many contexts. In comparison, as illustrated in Figure 1 (bottom right), a specific classifier has higher information density—specific classifiers are less frequent than the general classifier and typically lower-predictability—but, crucially, it constrains the hypothesis space for the identity of the upcoming noun, since the noun's referent must meet certain semantic requirement that the classifier is associated with. The UID hypothesis predicts that speakers choose a **specific** classifier more often when the predictability of the noun would other-

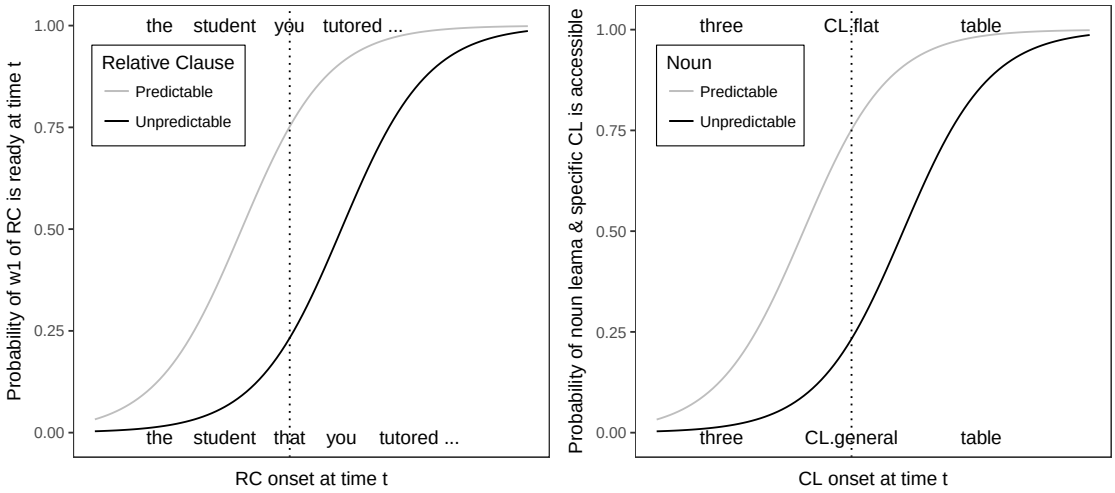

Figure 2: Schematic illustrations of availability-based production in the context of relative clause (left) and classifier choice (right). X axis presents the progression of time. The dotted lines indicate onset times for relative clause and classifier respectively.

wise be low.

Availability-based production, provided three plausible assumptions, makes different predictions than UID. The first assumption is that a speaker must access a noun lemma in order to access its appropriate specific classifier. The second assumption is that unpredictable noun lemmas are harder and/or slower to access (as described in Section 2, this assumption is supported by findings from the naming literature). The third assumption is that the general classifier is *always* available, regardless of the identity of the upcoming noun, as it is compatible with virtually every noun. Under these assumptions, for unpredictable nouns, specific classifiers will less often be available to the speaker when the time comes to initiate production of classifier, as shown in Figure 2 (right). Since noun lemmas need to be accessed before their associated specific classifiers, the less predictable the noun, the less likely the noun lemma and hence the associated specific classifier is to be available by the classifier onset time t. The general classifier, in contrast, is always accessible. Under these assumptions, the availability-based production hypothesis thus predicts that speakers choose a **general** classifier more often when the following noun is less predictable.

4 Data and Processing

To provide data for this study, we created a corpus of naturally occurring classifier-noun pairs from SogouCS, a collection of online news texts from various channels of Sohu News (Sogou, 2008). The deduplicated version of the corpus (see Section 4.1 for deduplication details) has 11,548,866 sentences. To parse and annotate the data, we built a pipeline to 1) clean and deduplicate the data, 2) part-of-speech tag and syntactically parse the clean text, and 3) extract and filter classifier-noun pairs from the parsed text. We are aware that a spoken corpus would be ideal to investigate speaker choice, but nothing this big is available. Instead we used SogouCS to approximate the language use of native speakers.

4.1 Cleaning and deduplication

Since the data contain web pages, many snippets are not meaningful content but automatically generated text such as legal notices. To use this corpus as a reasonable approximation of language experience of speakers, we performed deduplication on the data, following similar practice adopted by other work dealing with web-based corpora (Buck et al., 2014). After cleaning the text, we removed repeated lines in the corpus.

4.2 Word segmentation, POS-tagging and syntactic parsing

We used the Stanford CoreNLP toolkit for word segmentation, part-of-speech tagging, and syntactic parsing (Manning et al., 2014). We used CoreNLP's Shift-Reduce model for parsing (Zhu et al., 2013). We also got dependency parsing results as part of the Stanford CoreNLP output.

Noun	个 (ge, CL.general)	项 (xiang, CL.item)	张 (zhang, CL.flat)
公告 announcement	一 口 气 发布 11 个 公告 a CL breath release 11 CL announcement “release 11 announcements at one go”	连续 发布 三 项 公告 consecutively release three CL announcement “release three announcements in a” row”	门口 贴了 一 张 公告 door paste a CL announcement “there is an announcement on the door”
账单 bill	女儿 拿着 一 个 账单 就 过来了 daughter carry a CL bill at once come “daughter came with a bill at once”	not co-occurring	在 一 张 账单 上 解决 所有 收费 问题 on a CL bill solve all charge problem “solve all charge problems on a bill”
工程 project	跟 圆明园 有关 的 一 个 工程 to Yuanmingyuan related de a CL project “a project related to Yuanmingyuan”	抓好 六 项 重点 工程 grasp six CL key project “manage six key projects”	not co-occurring
活动 activity	昨天 我 参加了 一 个 活动 yesterday I attend a CL activity “yesterday I attended an activity”	广州市 今天 开展 的 一 项 活动 Guangzhou today hold de a CL activity “an activity held by Guangzhou today”	not co-occurring

Table 1: Examples from development set of available classifier options that are semantically (near-)invariant

4.3 Extracting and filtering classifier-noun pairs

From the parsed corpus, we extracted all observations where the head noun has a `nummod` relation with a numeral and the numeral has a `mark:clf` relation with a classifier. Figure 3 illustrates two such examples. We included classifiers in the list of 105 individual classifiers identified by Ma (2015) that are identified by the Stanford CoreNLP toolkit. For the purpose of restricting our data to cases of (nearly) semantically invariant alternation, we excluded classifiers such as *zhong* (“CL.kind”) that would introduce a clear truth-conditional change in utterance meaning, compared with the general classifier *ge*. We did further filtering to get nouns that can be used with both the general classifier and at least one specific classifier. This left us 1,479,579 observations of classifier-noun pairs.

To construct the development set, we randomly sampled about 10% of the noun types (1,179) and extracted all observations with of these noun types. We manually checked and filtered applicable classifiers for these noun types and we ended up with 713 noun types for the development set. For the test set, we also randomly sampled about 10% of the noun types (1,093) and extracted all observations with these noun types. We did not perform manual filtering of the test set. We reserve the remaining 80% for future work.

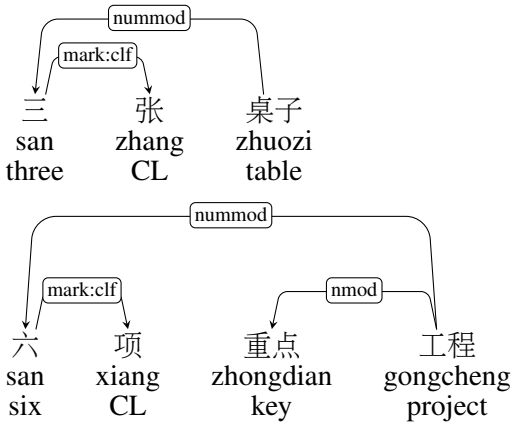

Figure 3: Classifier examples where the head noun has a nummod relation with a numeral and the numeral has a `mark:clf` relation with the classifier

5 Model estimation

We use SURPRISAL, the negative log probability of the word in the context (Hale, 2001; Levy, 2008; Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013), generated from a language model to estimate noun predictability. Since classifiers occur before their corresponding nouns, to avoid circularity, we mapped all target classifiers to the same token, `CL`, in the segmented text for language modeling, analogous to the procedure used in (Levy and Jaeger, 2007) and similar studies. We implemented 5gram modified Kneser-Ney smoothed models with the SRI Lan-

guage Modeling toolkit (Stolcke, 2002) and performed ten-fold cross-validation to estimate noun surprisal.

We used a mixed-effect logit model to investigate the relationship between noun predictability and classifier choice. The dependent variable was the binary outcome of whether a general or a specific classifier was used. For each noun type, we also identified its most frequently used specific classifier. We included two predictors in the analysis: noun surprisal and noun log frequency.[2] We included noun frequency as a control factor for two reasons. First, noun frequency has shown effects on many aspects of speaker behavior. Second, surprisal and frequency of a word are intrinsically correlated. Taken together, these two reasons make noun frequency an important potential confound to be controlled for in investigating any potential effect of noun surprisal on classifier choice.

We included noun and potential specific classifier as random factors, both with random intercepts and random slopes for noun surprisal. This random effect structure is maximal with regard to testing effects of noun surprisal, which varies within noun and within classifier (Barr et al., 2013). We then applied the model to the test set. The full formula in the style of `R`'s `lme4` package (Bates et al., 2014) is:

```
cl_choice~noun_surprisal+log_noun_freq
+(1+noun_surprisal|noun)
+(1+noun_surprisal|potential_spec_cl)
```

We used Markov chain Monte Carlo (MCMC) methods in the `R` package `MCMCglmm` (Hadfield et al., 2010) for significance testing, an based our p-values on the posterior distribution of regression model parameters using an uninformative prior and determining the largest possible symmetric posterior confidence interval on one side of zero, as is common for MCMC-based mixed model fitting (Baayen et al., 2008).

6 Results

In both the development set and the test set, overall we saw more observations with a specific classifier than with a general classifier (55.4% vs. 44.6% in the development set, 63.1% vs. 36.9% in the test set). For the development set, we find that the less predictable the noun, the less likely a specific classifier is to be used ($\beta = -0.038,\ p < 0.001$, Figure 4). There was no effect of noun frequency ($\beta = 0.018,\ p = 0.51$, Figure 5). For the test set, the result of noun predictability replicates ($\beta = -0.059,\ p < 0.001$, Figure 6).[3] In the test set but not in the development set, we also found an effect of noun frequency ($\beta = -0.11,\ p < 0.001$, Figure 7): the more frequent the noun, the less likely a specific classifier is to be used. Further analysis suggests that this effect of noun frequency in the test set is likely to be an artifact of incorrect noun–classifier associations that would disappear were we to filter the test set in the same way as we filtered the development set.[4] The consistent effect of noun surprisal on classifier choice in both our development and test sets supports the availability-based production hypothesis, and is inconsistent with the predictions of UID.

One potential concern regarding the above conclusion that noun predictability drives classifier choice is that it might not fully take into account effects of the frequencies of classifiers themselves on availability. The availability-based production hypothesis does not exclude the possibility that a classifier's accessibility is substantially dependent on its frequency, and the general classifier is indeed the most frequently used classifier. However, if specific classifier frequency were confounding the apparent effect of noun surprisal that we see in our analysis, there would have to be a correlation in our dataset between specific classifier frequency and noun surprisal. Our inclusion of a by-specific-classifier random intercept largely rules out the possibility that even a correlation that the above-mentioned one could be driving our effect. To be thorough, we tried a version of our regression analysis that also include a fixed effect for the log frequency of potential specific classifier as a control. We did not find any qualitative change to

[2] We used base 2 here to be consistent with the base used in noun surprisal.

[3] As can be seen in Figure 6, there is a bump at bin 27 in the rate of using a specific classifier. We consider this likely to be due to data sparsity: the number of observations is small in the last two bins of noun surprisal ($n = 27$ and $n = 3$), and there is no such bump in the development set.

[4] We found a marginal effect of noun frequency in our unfiltered development set, where the more frequent the noun was, the less likely it was used with a specific classifier. We did further analysis with the dev set and found that the"nouns" (some of them were misclassified as nouns from the results of the automatic parsing) that were excluded tend to have a higher frequency compared to the ones that were included, and the excluded ones also had a lower rate of concurring with a specific classifier. This tendency suggests that in the unfiltered test set, illegible nouns may contribute at least partially to the noun frequency effect.

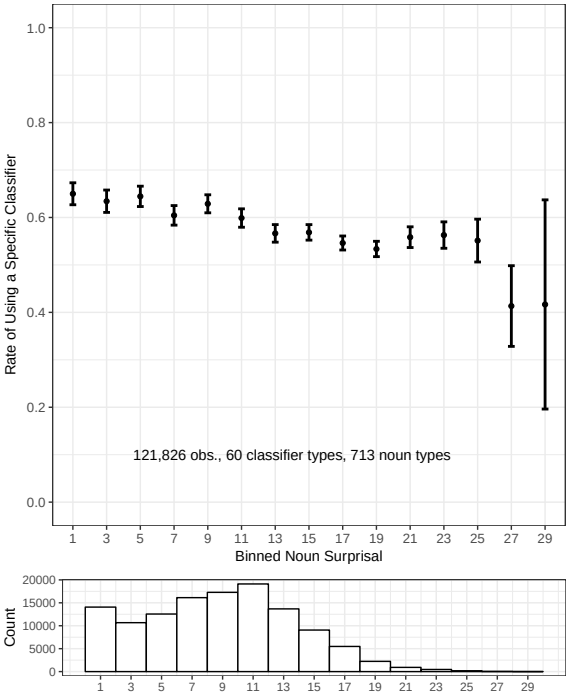

Figure 4: Dev set: N-gram estimated noun surprisal and the rate of using a specific classifier (as opposed to the general classifier *ge*).

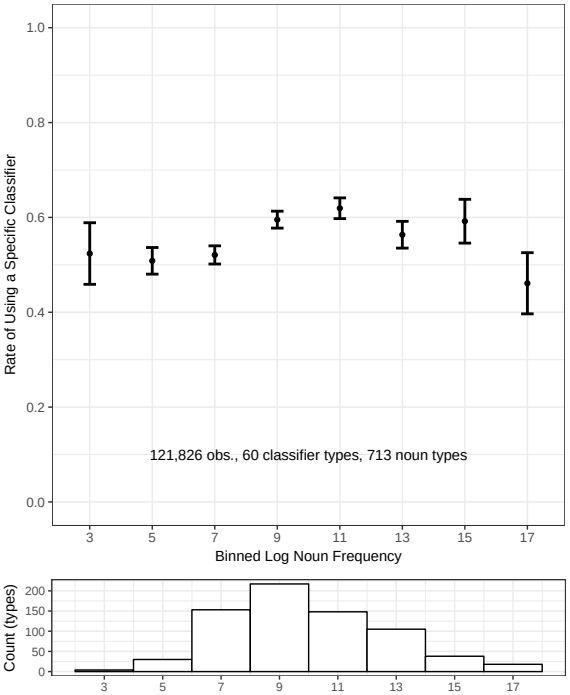

Figure 5: Dev set: Noun frequency (log scale) and the rate of using a specific classifier (as opposed to the general classifier *ge*).

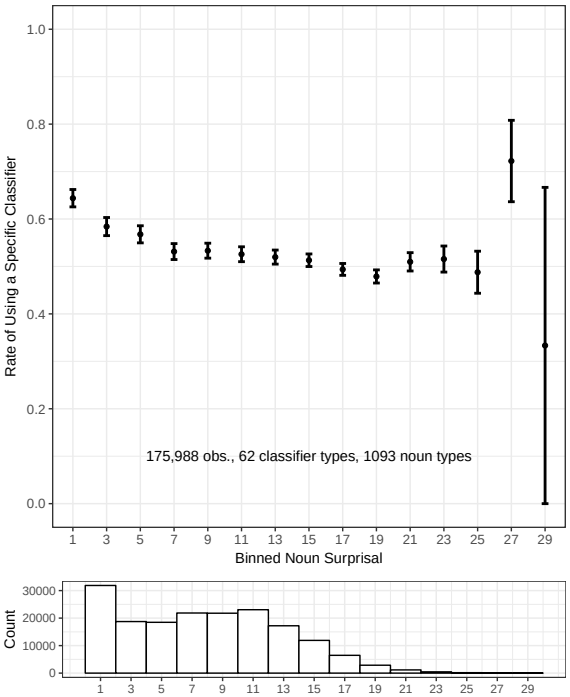

Figure 6: Test set: N-gram estimated noun surprisal and the rate of using a specific classifier.

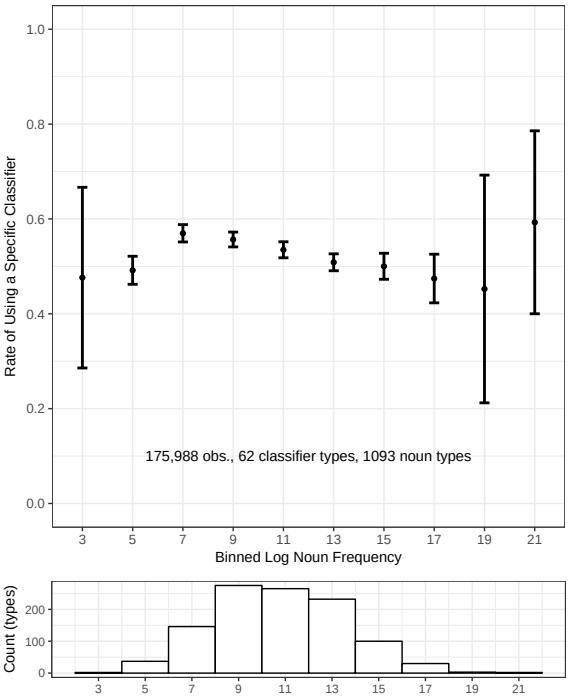

Figure 7: Test set: Noun frequency (log scale) and the rate of using a specific classifier.

the results: the effect of noun surprisal on specific classifier choice remains the same. We also note that in this new analysis, we do not find a significant effect of specific classifier log frequency on classifier choice ($p = 0.629$ for the dev set and $p = 0.7$ for the test set). This additional analysis suggests that it is unlikely that the effect of specific classifier frequency to be driving the effect of noun surprisal.

Overall, we did not find evidence for the UID hypothesis at the level of alternating options with different information density, in our case, a specific classifier versus a general classifier. We demonstrate that within the scope of near semantically invariant alternation, classifier choice is modulated by noun predictability with the tendency to facilitate speaker production. Our results lend support to an availability-based production model. We did not find consistent evidence for the effect of noun frequency on classifier choice. The effect of noun frequency remains unclear and we will need to test it with a larger sample of noun types.

7 Conclusion

Though it has proven difficult to disentangle UID and availability-based production through optional word omission phenomena, we have demonstrated here that the two accounts can potentially be distinguished through at least one word alternation phenomenon. The UID hypothesis predicts that predictable nouns favor the *general* classifier whereas availability-based production predicts that predictable nouns favor a *specific* classifier. Our empirical results favor the availability-based production account.

To the best of our knowledge, this is the first study that demonstrates contextual predictability is correlated with classifier choice. This study provides a starting point to understand the cognitive mechanisms governing speaker choices as manifested in various language optionalities. Ultimately we plan to complement our corpus analysis with real-time language production experiments to more throughly test hypotheses about speaker choice.

Acknowledgments

We gratefully acknowledge valuable feedback from Naomi Feldman, members of MIT's Computational Psycholinguistics Laboratory, three anonymous reviewers, technical advice for data processing from Wenzhe Qiu, and support from NSF grants BCS-1456081 and BCS-1551866 to RPL, and an MIT Henry E. Singleton (1940) Fellowship to MZ.

References

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4):390–412.

David A Balota and James I Chumbley. 1985. The locus of word-frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language* 24(1):89–106.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3):255–278.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. 2014. lme4: Linear mixed-effects models using eigen and s4. *R package version* 1(7).

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1):92–111.

Kathryn Bock. 1987. An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language* 26(2):119–137.

Christian Buck, Kenneth Heafield, and Bas Van Ooyen. 2014. N-gram counts and language models from the common crawl. In *LREC*. volume 2, page 4.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2):193–210.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/.

Victor S Ferreira and Gary S Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40(4):296–340.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science* 22(6):829–834.

Joseph H Greenberg. 1972. Numeral classifiers and substantival number: Problems in the genesis of a linguistic type. *Working Papers on Language Universals* 9.

Jarrod D Hadfield et al. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software* 33(2):1–22.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, pages 1–8. http://aclweb.org/anthology/N/N01/N01-1021.pdf.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1):23–62.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language* 45:229–254.

Natalie M Klein, Greg N Carlson, Renjie Li, T Florian Jaeger, and Michael K Tanenhaus. 2012. Classifying and massifying incrementally in chinese language comprehension. *Count and mass across languages* pages 261–282.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177.

Roger P. Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*. pages 849–856.

Aimin Ma. 2015. *Hanyu geti liangci de chansheng yu fazhan[The Emergence and Development of Chinese Individual Classifiers]*. China Social Sciences Press.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)*. pages 55–60. http://www.aclweb.org/anthology/P14-5010.

R.C. Oldfield and A. Wingfield. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17(4):273–281.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9):3526–3529.

Scott Seyfarth. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133(1):140–155.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3):302–319.

Sogou. 2008. Sogou lab data: Sohu news corpus 2008 version. http://www.sogou.com/labs/resource/cs.php. Accessed: 2017-05-30.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Interspeech*. volume 2002, page 2002.

Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and accurate shift-reduce constituent parsing. In *ACL (1)*. pages 434–443. http://www.aclweb.org/anthology/P13-1043.