Refer efficiently: Use less informative expressions for more predictable meanings

Harry Tily (hjt@stanford.edu)
Stanford Linguistics, Building 460, Serra Mall
Stanford, CA 94305 USA

Steven Piantadosi (piantado@mit.edu)
MIT Brain and Cognitive Sciences, 43 Vassar St
Cambridge, MA 02139 USA

Abstract

We present the results of a large-scale web experiment investigating comprehenders' ability to guess upcoming referents in an unfolding discourse. Participants were given a text that had been cut off just before a noun phrase, and attempted to guess which previously mentioned referent, if any, would be mentioned next. Our results show that writers are more likely to refer using a pronoun or proper name rather than a full NP when comprehenders have less uncertainty about the upcoming referent, and are more likely to use names than pronouns when comprehenders all tend to makes guesses to one or a few incorrect referents. These effects hold beyond other possible influences on the choice of referring expression type. Our results support addressee-oriented accounts of referring form choice (e.g. Brennan & Clark, 1996; Arnold, 2008) and suggest that language is a rational solution to the problem of communication: shorter and less informative expressions are favoured when less information is sufficient to carry the message (e.g. Jaeger, 2006; van Son & Pols, 2003).

Keywords: reference; pronominality; proper names; entropy; information theory; Shannon game

Introduction

Languages offer the possibility of producing arbitrarily specific noun phrases to pick out any single individual ("the President of the United States of America'). Many referents also have specific and unambiguous proper names ("Mr. Obama"). Yet, speakers often choose to refer to individuals and objects not by the most specific identifying phrase, but using linguistic forms which are non-specific and potentially ambiguous ("he"). The meaning of pronouns, in particular, depends on the current discourse context, and in isolation they are uninformative about the intended referent.

Intuitively, the advantage of using pronouns is that they are short and perhaps also easier to produce for other reasons (e.g. Ariel, 2001; Gordon, Grosz, & Gilliom, 1993; Almor, 1999). But their use tends to be reserved for situations where the intended referent is salient (e.g. Givon, 1976; Ariel, 2001; Gundel, Hedberg, & Zacharski, 1993; Almor & Nair, 2007; Arnold, 2008), avoiding confusion. Thus, pronouns save time and effort: they can convey the same message with less material. Zipf (1949) described a general tendency for brevity of expression, which he called the "Principle of Least Effort", and similar predictions follow from Gricean pragmatics. Using information theory, we can formalize these predictions. From Shannon's source coding theorem, the average utterance length will be minimized when the length of each word is proportional to the information it carries: the negative log probability with which the meaning could be guessed by the comprehender *before* hearing the word. Therefore, more predictable meanings should be given shorter words. As discussed above, referents may be more or less predictable depending on context, and so an optimal codelength can only be achieved if a single referent can be referred to with a longer or shorter expression depending on context. Pronouns in particular may allow for information-theoretically efficient communication, letting short phonetic forms be re-used for multiple meanings exactly when the omitted information is inferable from context.

Here, we directly test whether pronouns are used for referents which are more predictable in context. Shannon (1951) developed a method to measure the uncertainty (entropy) in language by measuring people's ability to guess the next letter or word of a text. More recently, Manin (2006) carried out a web-based experiment where people guessed the identity of a single word deleted from the middle of a text. His results showed that participants' accuracy was negatively correlated with the length of the word, supporting the view that language is an efficient code for meaning. However, in these games people guessed the upcoming linguistic units *themselves*, rather than their *meaning*. We present results from a similar "game", but one which measures uncertainty about meaning directly by measuring people's ability to guess what established discourse referent the next upcoming NP refers to.

Experiments

We prepared a database of 82 texts extracted from the Wall Street Journal section of the OntoNotes corpus (Weischedel, Pradhan, Ramshaw, & Micciulla, 2008). Texts were truncated after the 30th noun phrase if they contained more than 30, yielding a total of 2211 noun phrases. Texts were presented piece by piece to participants on the Amazon Mechanical Turk website (`www.mturk.com`). The first piece of text included the text from the start of the document up until immediately before the first NP; the second piece included the first NP and all text immediately up until just before the second, and so on. Participants were told they were playing a "guessing game", and their job after seeing each piece was to guess what would be mentioned next. They would click on a previous NP in the text if they expected that previous referent to be mentioned again, or on a "something new" button if they thought a new referent would be introduced. After each click, the next section was revealed and participants invited to guess again. NPs were coloured such that all NPs with a common reference appeared in the same shade. We stressed that the task was not to predict the particular *words* that would be used, but the *person or thing* that would be referred to. Participants received feedback telling them whether they guessed correctly, and a running score showing their overall performance. Since most NPs in the texts were only mentioned once (75%), in most trials the correct answer was “something new”. To encourage participants not to choose that response on every trial, we gave 1 point for a correct “something new” response and 2 points for correct clicks to a previous NP. Additionally, on each trial the “something new” button appeared in a slightly different location so that participants could not repeatedly click on it without moving the mouse.

471 people (all from the USA) participated, with nearly 50 people seeing each text. We excluded all trials from participants who returned partial submissions, saw less than 3 documents or clicked one response throughout. We also excluded trials submitted very quickly (below the 5th percentile). In total, 21.4% of the data was excluded in this way, leaving 71644 noun phrase trials.

We coded each NP for whether it was a pronoun or proper name; along with “description”, the remainder of NPs, this yields a three-way distinction between referring expression types. We also coded each NP for several other properties that we thought might predict the writer’s choice of referring expression or the participant’s ability to guess:

- *NP number*: number of the NP in the discourse
- *sentence number*: number of the sentence in the discourse
- *referents*: total number of discourse referents introduced up to this point
- *mentions*: number of times this referent has been mentioned so far (discourse saliency)
- *distance to last mention*: measured in intervening words, and NPs
- *grammatical function*: whether the NP is (a) a subject, (b) a direct object, or (c) neither
- *previous subject*: whether this NP is coreferential with the subject of the previous clause
- *previous object*: whether this NP is coreferential with the direct object of the previous clause
- *previous expression type*: whether the previous NP coreferential with this one was a pronoun, a proper name, or other

We only consider *subsequent-mention* trials, those where the correct referent has already been mentioned (25%, 18227 trials). The correct response for first-mention trials is “something new”, making it impossible to estimate the probability a participant was expecting the specific correct referent.

When can comprehenders predict reference?

As a first investigation, we tested whether comprehenders are able to predict which referent the writer will next refer to more or less accurately depending on the type of referring expression. As described above, this guess is made *before* the referring expression is seen, so any relationship between the two variables will suggest that the choice of referring expression type is influenced by predictability, not that different referring expression forms influence guessing accuracy.

Using multilevel logistic regression, we regressed whether participants chose the correct referent on each trial against the control variables described above. As a correct response would sometimes require clicking on one particular short previous NP, and sometimes on any of several long NPs, we also included a control coding the total screen area taken up by correct responses, to control for random clicking. We included random intercepts for participant, document, and NP identity. Our independent variable of interest was the three-way categorical variable coding referring expression type: pronoun, name or description. We arrived at a final model using the “drop 1” procedure, eliminating predictors one at a time that did not improve goodness of fit to the data by chi-square model comparison.

The results (in Table 1) show the differences in participants’ ability to guess the reference depending on each of the predictors. Positive coefficients indicate a higher probability of guessing, and negative a lower probability. Continuous predictors are standardised following the procedure in Gelman (2008), meaning that the size of coefficients for each predictor can be approximately compared to give an indication of how important that predictor is relative to the others. We report the chi-square value associated with the reduction in data log likelihood when removing each predictor, which was the basis of the model comparison used to arrive at this final model. The model was fitted using the lme4 package for R (Bates & Maechler, 2009). Since degrees of freedom cannot be straightforwardly determined for multilevel regression models, coefficient p values were obtained through Monte Carlo simulation using the function `pvals.fnc` in the R package languageR (Baayen, 2008).

The model summary in Table 1 shows the expected effects: participants guess correctly more often when the upcoming referring expression is a pronoun, even before they have seen the pronoun, and even though the task is not to guess the form of the expression but the referent itself. Accuracy is equally high with an upcoming proper name, which is not surprising since referents that have proper names are often more salient and topical.[1] Referents which have been mentioned more times before are guessed more often, leading to higher accuracy. Also, as the number of referents in the discourse increases, accuracy decreases, presumably due to competition. Although there was no detectable difference in comprehenders’ ability to guess reference for direct objects and more oblique uses, grammatical subjects were guessed *less* accurately than more oblique NPs, controlling for other factors.

[1] If we could look just at the subset of referents that *could* conceivably have been referred to by proper name instead of those that *were*, we would not necessarily expect to see this effect.

		β	s.e.	z	p_z	χ^2	df	p_{χ^2}
Intercept		-3.2	.17	-18	<2.0e-16			
expression type	pronoun	.55	.17	3.2	.0015	13.14	2	.0014
	name	.56	.17	3.2	.0015			
number of mentions		.38	.058	6.6	4.3e-11	41.86	1	9.8e-11
number of referents		-.64	.075	-8.5	<2.0e-16	67.82	1	2.2e-16
previous type	pronoun	.43	.16	2.7	.0068	65.69	2	5.45e-15
	name	.84	.19	4.5	6.0e-06			
grammatical function	subject	-.61	.17	-3.5	.00042	20.69	2	3.21e-05
	object	.23	.24	.97	.33			

Table 1: Logistic model, predicting a correct guess for each trial

This may simply be because subjects tend to occur at the start of the sentence, at which point there is relatively less information available. Finally, guessing was more accurate for NPs which had last been referred to with a pronoun than a description, and more accurate still for those last referred to with a name. Other fixed effect controls were not significant by model comparison. Interestingly, these include distance from last mention, which perhaps indicates that salience is mediated by other predictors (previous expression type and number of mentions). Model comparison tests of the random intercepts showed significant variation between comprehenders, documents, and individual NPs.

Although uncertainty in comprehension and choice of expression in production are apparently connected, the referring form cannot influence the participants' ability to guess, since guessing takes place before the expression is revealed. We adhere to the opposite causal story: writers use less informative expressions when comprehenders are better able to guess the referent.

When do writers use different forms?

To test whether writers' choice between using pronouns, names, and descriptions is influenced by the predictability of the referent, we performed a second analysis of the data. To claim that predictability influences referring expression choice, we need to model that choice directly as an outcome, and show that predictability matters even after controlling for many other factors that may also influence the choice.

We used a multinomial logit model in which datapoints each represented one NP, and the outcome variable was the referring expression type used: pronoun, name, or description. We included as predictors all of the variables described above as well as two measures of interest which gauge comprehender's ability to guess the correct referent for a given NP. The first of these is the negative log probability of a correct guess, $-\log_2 P(Response = correct)$. This is a direct measure of the *information* conveyed by the NP: if comprehenders always guess correctly without needing the word, it carries zero information. This value is commonly called *surprisal* in the psycholinguistic literature. However, surprisal reduces the guess to a binary outcome (correct or incorrect) ignoring potential differences in the distribution over guesses. For instance, two NPs may both be guessed correctly 50% of the time, but one may have a high-probability competitor referent which is incorrectly guessed the remaining 50% of the time, while guesses in the other case are evenly spread over a large number of low-probability referents. This could plausibly influence a writer's choice of referring form: for example, "he" might be chosen less often when a second male referent is also highly salient/activated, even controlling for the ease of guessing the intended referent. To capture this, we also included a measure of the *entropy* of the distribution, $H(Response) = -\sum_{r \in Response} P(r) \log P(r)$. This value will be low when (correct or incorrect) guesses are concentrated on one referent and high when they are spread over many.[2] Naturally, these two measures are closely related, as shown in Figure 1. To avoid collinearity problems, we used residuals of the entropy measure as plotted on the right of that figure.[3] The conditional density plots suggest a relatively weak relationship between surprisal and referring expression form, with fewer pronouns and names and more descriptions as guessing accuracy decreases. Conversely, the ratio of descriptions to names does not vary greatly with entropy, but high entropy situations seem to strongly favour pronouns.[4]

The model was fitted using the mlogit procedure in the package Zelig (Imai, King, & Lau, 2007), and the final model chosen using drop 1 chi-square model comparison. A multinomial logit model with a three-way response variable chooses one level as a baseline (here "Description"), equating it with a value of zero. Coefficients are then estimated for each predictor for each of the other two levels, so that each datapoint is associated with two values, obtained by taking the product of the vector of data and each of the two coefficient vectors. These values can then be turned into predic-

[2] In calculating entropy, we assume that each "something else" response refers to a distinct referent, which is a simplification but preferable to assuming that all those answers are guesses to the *same* referent.

[3] Residuals were taken on $P(correct)$ rather than its log, since the relationship between those variables is extremely close to linear. After residualising entropy and centering all continuous predictors, collinearity was not a problem; the condition number of the matrix of predictors in the final model is 12.66.

[4] These relationships hold at least for the central region of the density plots, where most of the data lies.

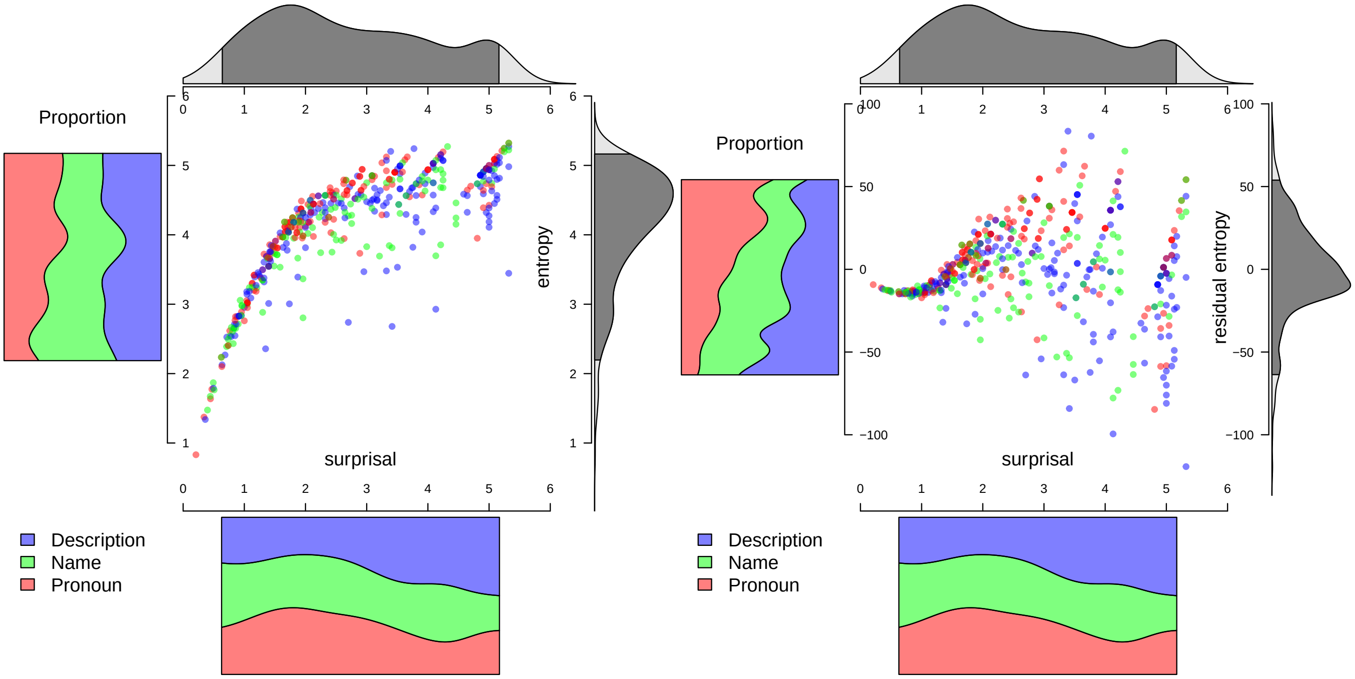

Figure 1: *(left)* The relationship between comprehenders' ability to guess the specific upcoming referent (surprisal) and global uncertainty about the upcoming referent (entropy). NPs that more people guessed correctly appear further to the left and those where guesses were more evenly distributed over many answers appear further up. Points in the bottom left are those where all participants systematically guessed the correct answer. Density plots in the top and right show the marginal distribution of the two variables, and the conditional density plots in the bottom and left show the proportion of NPs of each referring type as a function of each variable. *(right)* The same information using standardised residuals of entropy after decorrelating it from the probability of a correct guess.

tions in probability space: estimates of the probability that each NP is a pronoun, name, or description, with those probabilities constrained to sum to 1. To do this, we take the exponent of the two values associated with the two non-baseline levels ("Pronoun" and "Name") and the exponent of the baseline level (1) and normalise these three values by dividing by their sum. This means that just like in a logistic regression, positive coefficients for one outcome lead to a higher probability of that outcome relative to the baseline level, and negative coefficients lower. Likewise, if the first outcome level has a coefficient that is greater than that of the second outcome, higher values of the associated predictor lead to more probability of the first outcome relative to the second.

The final model is displayed in Table 2. Additionally, the model's predictions for the dataset are plotted in Figure 2. These predictions are displayed in a "ternary plot": a triangle where the corners correspond to probability 1 for one outcome and 0 for the other two, and the centre corresponds to probability 1/3 for each. The model correctly assigns high pronoun pronoun to most pronouns and high description probability to most descriptions, but seems to do rather less well at detecting names.

Several strong effects are evident from the table of coefficients. For instance, later NPs are much less likely to be pronouns than descriptions after controlling for all other factors. So are referents that were last referred to a longer distance ago in the text. The form of the previous coreferential expression greatly influences the choice between types, with both pronouns and names more likely if the previous mention was also a pronoun or name, although the tendency to reuse pronouns is much weaker than that to reuse names — perhaps due to stylistic avoidance of repeated pronoun mentions. Pronouns are very likely if the referent was the subject or object of the previous clause, and names slightly disprefered relative to descriptions or pronouns. When the NP is a subject, pronouns are strongly prefered over descriptions and somewhat over names. For direct objects, pronouns are again prefered, but descriptions and names are roughly equally likely. Turning to the two variables of most interest, both pronouns and names are disfavoured relative to descriptions when comprehenders have difficulty guessing the correct referent, and they are disfavoured to a fairly similar degree. On the other hand, entropy does not discriminate as well between names and descriptions, but pronouns are favoured over the other two types as entropy increases. Both entropy and surprisal interact with the previous expression type, in particular making names following pronouns less likely when entropy is high, and more likely when surpisal is high.

The influence of each predictor on the model can be visualised by plotting the change in the model's predictions

		Pronoun			Name					
		β	s.e.	*t*	β	s.e.	*t*	χ^2	df	p_{χ^2}
Intercept		-4.7	1.0	-4.5	-1.4	.34	-4.0			
sentence no		-.044	1.0	-.044	1.3	.41	3.1	31	8	.00014
NP no		-2.2	.66	-3.3	-.15	.54	-.28	24	4	9.1e-5
sentence no ∗ NP no		.031	.19	.17	-.48	.19	-2.5	8.7	2	.013
no referents		1.6	.59	2.7	.11	.47	.24	8.5	2	.014
previous type	pronoun	1.2	.39	3.0	2.0	.33	6.1	71	12	1.9e-10
	name	.81	.45	1.8	.58	.51	1.2			
last mention distance		-1.4	.30	-4.6	-.00094	.16	-.0060	34	2	5.3e-8
subject of last clause		1.7	.43	4.1	-.12	.51	-.24	25	2	4.2e-6
object of last clause		1.6	.69	2.3	-.066	.77	-.087	7.4	2	.025
grammatical function	subject	3.7	1.0	3.6	.77	.34	2.2	58	8	1.4e-9
	object	1.9	1.1	1.7	-.17	.49	-.34			
sentence no ∗ grammatical function	subject	1.1	.98	1.1	-.74	.38	-1.9	12	4	.018
	object	.27	1.1	.25	-1.3	.56	-2.3			
residual entropy		.76	.25	3.1	.35	.23	1.5	23	6	.00087
surprisal		-.43	.26	-1.7	-.52	.23	-2.3	15	6	.019
entropy ∗ previous type	pronoun	-.65	.41	-1.6	-1.0	.37	-2.7	9.4	4	.051
	name	-.75	.47	-1.6	-.32	.54	-.59			
surprisal ∗ previous type	pronoun	.24	.40	.61	.73	.35	2.1	11	4	.030
	name	.12	.44	.26	-.53	.53	-1.0			

Table 2: Multinomial logit model predicting referring expression type for each NP (baseline level is "Description"). Chi-square values are for removal of a predictor *and* any higher-order interactions it participates in.

for the data as predictors are taken into consideration. Figure 3 shows how the model's estimates for particular NPs change when predictors are removed from the final model. Adding any predictor should have the effect of pushing the data "closer to the corners" on average, since adding predictors always makes the data more likely. However, predictors differ in the specific datapoints they affect. For instance, both the subject/object of previous clause predictors and the lastmention predictors act by making pronouns more or less likely relative to the other outcomes, since all movement is into or out of the bottom left corner. However, while knowing about the previous subject/object biases the model strongly towards guessing pronouns, apparently increasing the rate of false pronoun guesses, knowing the distance from the last mention seems to let the model discriminate more accurately between pronoun and nonpronoun choices, correctly assigning more probability of pronominality to actual pronoun outcomes and less to actual nonpronoun outcomes. The surprisal and entropy predictors seem to help discrimination between all three categories, since there is movement parallel with all three edges of the plot. However, it can be seen that the largest effect on the prediction is in more accurately classifying descriptions as such (and particularly, descriptions that would otherwise have been categorised as pronouns).

Discussion

Our results investigate the relationship between the choice of referring expression and the predictability of the referent it picks out. Importantly, these results are *not* about the difficult of resolving a pronoun itself – such a study would measure the difficulty of guessing what each NP refers to. Instead, we measured how well subjects could guess what the next upcoming NP would refer to, without seeing its linguistic realization.

Our first study investigated the factors that contribute to a comprehender's ability to guess an upcoming referent predictability. We found that referents are more predictable when they are are mentioned more often, occur with fewer other referents in the discourse, or were referred to most recently with a pronoun or name. These findings indicate that comprehenders construct sophisticated models of discourse for resolving and predicting reference, and are sensitive to several types of cue in doing so.

Our second study empirically tested the theory that pronouns allow for more concise communication by allowing speakers to refer to highly predictable referents with short words. The analysis showed that pronouns are used in places where comprehenders can more easily guess the upcoming referent, suggesting that language users choose referring expressions appropriate to the comprehender's uncertainty. Even so, the relationship between referring expressions and uncertainty is nuanced. Above and beyond the effect of guess accuracy on expression choice, there is also an effect of guess *entropy*. If a comprehender is unable to guess the upcoming referent, then writers tend to use descriptions. However, if the comprehender has a higher probability of guessing,

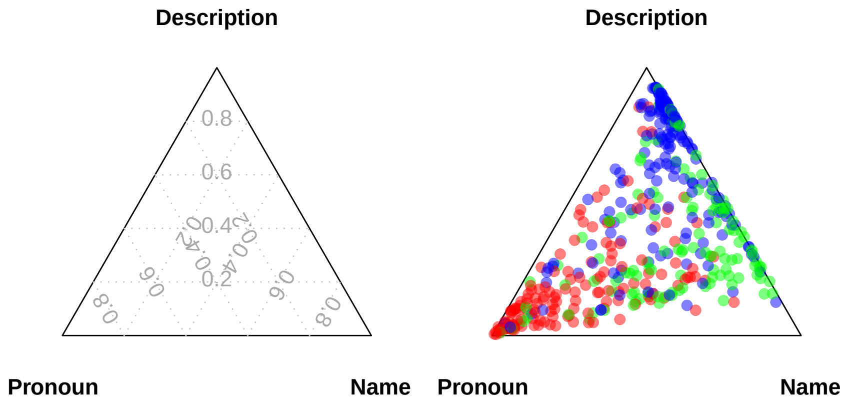

Figure 2: *(left)* A guide for interpretation of ternary probability plots. Each point corresponds to a different division of probability between the three outcomes, with points along one sides representing probability 0 for one outcome, and lines parallel to that side being contours of equal probability for that outcome. *(right)* The model's final predictions on the data set. An ideal model would cluster all true pronouns (red) in the bottom left, true names (green) in the bottom right, and true descriptions (blue) at the top.

then the choice between pronoun and name depends partly on the presence of competing discourse referents: if incorrect guesses are concentrated on one or a few high-probability competitors, names are preferred, while if the intended referent is the only potential referent with high probability, then pronouns are preferred.

Beyond the two uncertainty-based predictors, our second model showed that pronouns tend to be used when they are recently mentioned (particularly as a previous subject or direct object), in referent-heavy discourses, coreferential with previous names or pronouns, and when they are themselves subjects or direct objects. Proper names, on the other hand, are favoured for referents that were last referred to with a name or pronoun, and slightly favoured over descriptions for grammatical subjects. As both these discourse internal measures *and* comprehender uncertainty measures proved significant, we do not believe that either can be reduced to the other: the choice of expression type shows effects of both addressee-oriented design (e.g. Brennan & Clark, 1996; Arnold, 1998, 2008) and speaker-internal factors; perhaps memory or accessibility, perhaps simply stylistics.

In the field of Generating Referring Expressions (GRE), and particularly the work of Dale & Reiter (e.g. Dale & Reiter (1995)) an emphasis has been placed on choosing an expression that not only picks out the intended referent unambiguously, but also accords with Gricean maxims of Quantity and Brevity: a GRE system should choose an expression that is as brief as possible and introduces as little information as possible beyond what is needed to pick out the referent. Our results confirm that human expression choice is appropriate to the comprehender's level of uncertainty, at least in the choice between pronouns, names, and descriptions. More recent GRE models try to maximize brevity while discriminating only between the most salient discourse referents (Kraemer & Theune, 2004), investigating new ways to determine the relevant *context set* of referents that must be discriminated between. The online methodology we introduced could be used to determine empirically the context set that comprehenders entertain, which could lead to applications in GRE given suitably annotated data.

In some ways, the task here is not ideal for measuring the information conveyed by a noun phrase. We have made the tacit assumption that the log probability of a correct guess is equal to the information that will be conveyed by the NP. In fact, there is no guarantee that the comprehender is entirely certain about the correct reference after processing the NP; therefore, uncertainty may be reduced but not entirely eliminated. This does not seem unrealistic given that in general, words can only be understood in combination with the other words in the discourse, and therefore the contribution of many words will only be fully understood once later words have also been processed. In fact, in some cases it may be advantageous to stretch out information over multiple words (Jaeger, 2006; Levy & Jaeger, 2007). An alternative task could have one NP replaced by a blank in the middle of a text, and collect guesses that would be informed by both previous and following discourse. It would be interesting to see which measure better correlates with the referring expression choice.

Another potential issue with these results is that they come from written data, specifically from newspaper text. As well as being somewhat unrepresentative of everyday language, written data is necessarily one-sided, being written in the absence of a comprehender. In audience-oriented theories such as Brennan & Clark (1996), the choice of referring expression is a "conceptual pact" which is mutual between two parties. We are currently re-running this experiment using conversational data, to examine whether comprehender uncertainty is more important when the comprehender is present and interacting in the discourse.

In other analyses of this data, we looked for but failed to find a relationship between surprisal/entropy and the *length* of the NP, beyond the pronoun/name/description distinction.

This might be taken as evidence against a general theory that shorter forms are preferred in predictable situations. However, the descriptions in this newspaper text include a large number of extremely long NPs, which are particular to the genre and would be odd in natural speech. We will return to this question when we have results from spoken data.

In summary, the empirical data analysed here is compatible with a view of language as an efficient code for communication. Pronouns in particular provide language with context-dependent code that allows more predictable nouns to be referenced with a shorter word. These results align with recent production theories such as Uniform Information Density (e.g. Jaeger, 2006; Levy & Jaeger, 2007; Genzel & Charniak, 2002).

Like much of language, pronouns are, in some sense, ambiguous. Simple NPs such as generic nouns ("the sailor"), names ("John"), and even more complex NPs are rarely, if ever, are completely unambiguous[5], and making them so would undoubtedly make language less concise and sound more like legalese. Pressure for efficiency may explain why language can tolerate such superficially unclear expressions. Efficient communication does not mean that each unit alone should completely convey the intended meaning, but that language, along with context, world knowledge, and human capacity for inference, should be sufficient to recover the meaning. Efficient language should be as concise as possible given these other informative cues, which entails using short ambiguous linguistic constructions like pronouns exactly when speakers can use other information to disambiguate.

References

Almor, A. (1999). Noun-phrase anaphora and focus: the informational load hypothesis. *Psychological Review*, *106*, 748-765.

Almor, A., & Nair, V. (2007). The form of referential expressions in discourse. *Language and Linguistics Compass*, 84-99.

Ariel, M. (2001). Accessibility theory: An overview. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (p. 29-87). Amsterdam: John Benjamins.

Arnold, J. (1998). *Reference form and discourse patterns*. Unpublished doctoral dissertation, Stanford University.

Arnold, J. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, *23*, 495-527.

Baayen, R. H. (2008). *Practical data analysis for the language sciences with R.* Cambridge, UK: Cambridge University Press.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes [Computer software manual]. Available from `http://CRAN.R-project.org/package=lme4` (R package version 0.999375-31)

Brennan, S., & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, *22*, 1482-1493.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, *18*, 233-263.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*, 2865-2873.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the ACL.* Philadelphia: UPenn.

Givon, T. (1976). Topic, pronoun and grammatical agreement. In C. Li (Ed.), *Subject and topic.* New York: Academic Press.

Gordon, P., Grosz, B., & Gilliom, L. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*, 311-47.

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language and Cognitive Processes*, *69*, 274-307.

Imai, K., King, G., & Lau, O. (2007). mlogit: Multinomial logistic regression for dependent variables with unordered categorical values [Computer software manual]. Available from `http://gking.harvard.edu/zelig`

Jaeger, T. F. (2006). *Probabilistic syntactic production: Expectedness and syntactic reduction in spontaneous speech.* Unpublished doctoral dissertation, Stanford University.

Kraemer, E., & Theune, M. (2004). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), *Information sharing.* Stanford: CSLI.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19.* Cambridge, MA: MIT Press.

Manin, D. (2006). Experiments on predictability of word in context and information rate in natural language. *Journal of Information Processes*, *6*, 229-236.

Shannon, C. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *3*, 50-64.

van Son, R., & Pols, L. (2003). How efficient is speech? *Proceedings of the Institute of Phonetic Sciences*, *25*, 171-184.

Weischedel, R., Pradhan, S., Ramshaw, L., & Micciulla, L. (2008). *Ontonotes release 2.0.* (Linguistic Data Consortium)

Zipf, G. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology.* New York: Hafner.

[5] For instance "the man in the park with the new running shoes" could refer to any man in any park who has new running shoes.

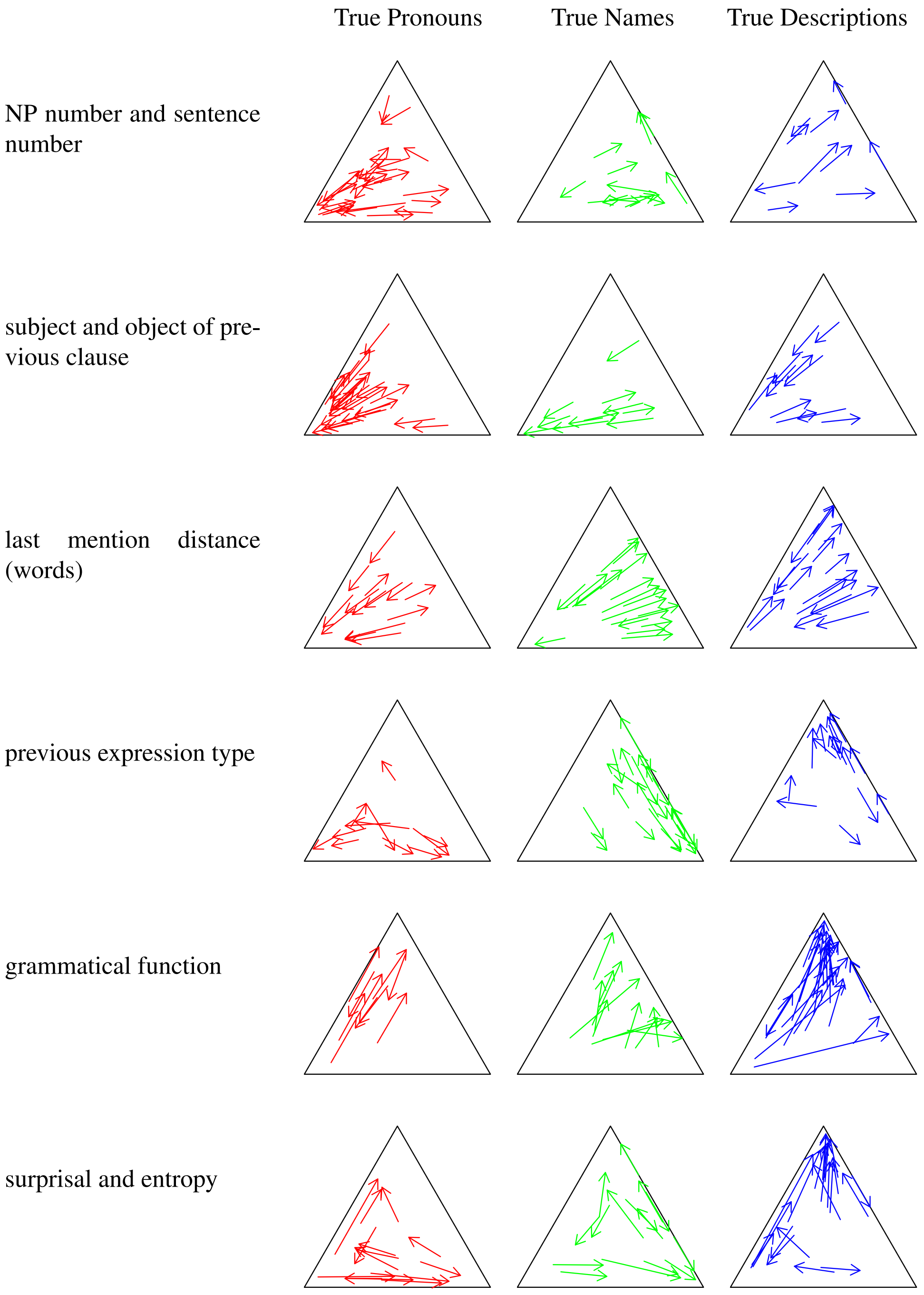

Figure 3: Changes in model predictions as sets of predictors are included. Arrows join the prediction of the model with all significant predictors *except* those indicated to the predictions of the final model for the same data point, for the 50 NPs whose predictions are most affected.