PROSODIC MARKING AND PREDICTABILITY IN LEXICAL SELF-REPAIR

Leendert Plug

University of Leeds
l.plug@leeds.ac.uk

ABSTRACT

This paper reports on an investigation of lexical self-repair in Dutch spontaneous dialogue. Lexical self-repairs, in which one word is rejected for another, can be produced with or without notable 'prosodic marking' of the second word. It remains unclear what motivates speakers' choices, but previous research has shown that the semantic distance between the two words is relevant. This study assesses the relevance of the words' predictability. Prosodic marking judgements are modelled using an established semantic classification and a range of probabilistic variables, including both frequency-based and cloze-based measures. Results suggest that probabilistic measures add little predictive power to the semantic classification, although informative data trends can be observed.

Keywords: prosody, self-repair, predictability, spontaneous speech, Dutch

1. INTRODUCTION

This paper reports on an investigation of lexical self-repair, in which a speaker rejects one lexical choice in favour of another: e.g. *on Thursd- Friday*. Previous research has shown that the correction can be produced with or without 'prosodic marking' [2, 3, 5, 6]. In an 'unmarked' production, the pitch, intensity and speaking rate of the repair word or phrase—here *Friday*—are similar to those of the reparandum—*Thursd-*. In a 'marked' production, the repair "is distinguished by a quite different prosodic shape from that of the original utterance" [2: 81]; this generally involves high pitch and intensity.

A relevant question is what motivates a speaker to produce a self-repair with prosodic marking. The literature on repair contains two proposals. [5] argues that speakers mark repairs to highlight particularly salient information, facilitating listeners' comprehension. [5] cites [6], which reports that repairs in which factual or linguistic errors are corrected are more often marked than repairs in which subtler 'appropriateness' issues are addressed. According to [5], error repairs are associated with greater semantic contrast between reparandum and repair than appropriateness repairs. The higher the degree of contrast, the more informative the repair, so the more motivation there is for marking. Taking a different tack, [3] suggests that speakers use marking to save face: according to [3], marking diverts listeners' attention away from a problematic formulation. Therefore, the more conspicuous and potentially embarrassing the reparandum, the more likely it is that a speaker will mark its correction.

Subsequent research has reported similar patterns to those in [6], but shown that semantically-based measures offer limited prediction of repair prosody [8, 10]. Of course, semantic distance measures can only partly capture the information value of a repair [5: 496]. Another major dimension is predictability. The reasoning in [5] predicts a negative relationship between marking and predictability: repairs with unpredictable lexical items should be marked more often than repairs with predictable, therefore 'informationally redundant' ones [1, 9, 11]. By contrast, the reasoning in [3] focuses attention on the predictability of the repair *per se*, and predicts a positive relationship with marking: formulations that are predictably in need of correction should be marked more often than those whose reparanda are not easily identified as problematic.

This study assesses the relevance of repair and repair component predictability for understanding the distribution of prosodic marking in self-repair. It does so by implementing frequency-based measures and cloze probabilities to estimate the predictability of the repair *per se*, as well as the repair component. Frequency-based probability estimates are common in corpus-based research, and measures based on n-gram frequencies offer a degree of context-sensitivity [9, 11, 12]. Still, they remain distinct from cloze probabilities elicited through fill-in-the-gap tasks, which fully reflect the contribution of prior discourse context and general world knowledge to predictability [7]. In this study, both types of measure are assessed as candidate predictors of prosodic marking, alongside a semantic classification based on [5, 6]. The crucial question is whether probabilistic measures add to, interact with or outperform the semantic classification.

2. DATA AND METHOD

The data for this paper comprise 209 instances of lexical self-repair extracted from sub-corpora of the Spoken Dutch Corpus containing spontaneous dialogue. They only include instances in which one lexical item is replaced by another. Instances with an incomplete reparandum item were included if a good guess could be made as to its identity. Examples include *met de au- met de bus* 'by ca- by bus' and *een leuke k- een mooie keuken* 'a nice k- a beautiful kitchen'. All instances are utterance-medial.

2.1. Prosodic analysis

Each instance was judged prosodically marked or unmarked by two Dutch linguists. Judging was done independently through auditory analysis, with the option of judging an instance 'possibly marked'. The data comprised 216 instances, 7 of which were later excluded (see below). The raters reached the same judgement for 182 (84%); for the remaining instances, a consensus was reached. Instances judged 'possibly marked' were recoded as 'marked' for this study, yielding 143 'unmarked' instances (68%) and 66 'marked' ones (32%).

2.2. Semantic classification

Each instance was classified as appropriateness or error repair, as in [5, 6]. Instances involving a factual inaccuracy or linguistic ill-formedness are error repairs; all others are appropriateness repairs. Classification was done independently by two Dutch linguists, for 222 instances. Their classifications matched for 201 (91%). A consensus was reached for 15; 6 were excluded. (A further 7 were later excluded: see below.) Error repairs were additionally coded 'factual' or 'linguistic', as in [10].

2.3. Frequency-based measures

For each instance, and for both lexical items in it, I took unigram word frequency counts from CELEX and bigram counts (with pre- and post-repair items) from the Spoken Dutch Corpus. Following [9, 12], we can expect the bigram counts to perform similarly to trigram counts or more complicated models. In addition to entering the (log-transformed, centered) counts into the analysis, I subtracted each reparandum count from the corresponding repair count to yield a measure of the relative predictability of the repair item.

2.4. Cloze probabilities

To obtain more context-sensitive measures of repair item and repair predictability, I devised two fill-in-the-gap tasks. First, I transcribed all instances in their phrasal context with the reparandum item present but the repair item withheld. Incomplete reparanda were completed and reparanda highlighted for clarity: e.g. *met de au- met de bus* was rendered *met de* ***auto*** *met de* ___. Prior discourse was summarised and any previous mentions of the repair item made explicit. Three native speakers of Dutch provided up to two candidate repair items, ranked as first and second choice. Responses were quantified according to whether the rater guessed the correct lexical item and offered it as first or second choice. The data comprised 216 instances. 7 were found to contain a transcription error or to be interpretable as a grammatical repair; these were excluded from the analysis. Responses for the remaining 209 yielded an acceptable ICC (0.75). The scores were summed to produce a scale of repair item predictability from 0 (not predictable) to 12 (highly predictable).

I also transcribed all instances using the same method, but with the entire repair withheld: *met de au- met de bus* was rendered *met de* ***auto***. Three native speakers were asked whether the highlighted word choice was in need of repair—unaware that *all* were in fact followed by repair. The raters' binary judgements were quantified yielding a very high ICC (0.93). Responses were summed to produce a scale of repair predictability from 0 (not predictable) to 3 (highly predictable).

3. RESULTS

3.1. Relationships among predictor variables

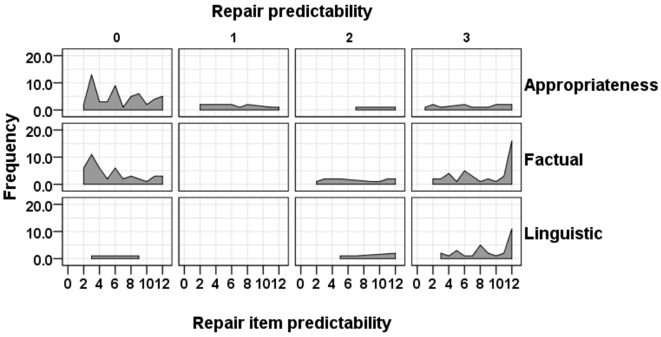
Before assessing the performance of the candidate predictors above in modelling prosodic marking, we can explore the relationships among them. As might be expected, unigram and bigram frequencies are significantly correlated with each other. The strongest correlation is that between the unigram frequencies of reparandum and repair items (Pearson's r=0.793, p<0.001). As found by [4], low-frequency repair items tend to be preceded by higher-frequency reparanda.

The cloze-based measures of repair item and repair predictability are also significantly correlated with each other (r=0.320, p<0.001). The correlation means that repairs whose reparandum is not clearly erroneous also tend to have repair items that are difficult to predict; repairs of easily identifiable errors also tend to have obvious resolutions. The cloze-based measure of repair predictability is not significantly correlated with any of the frequency measures. Thus, low-frequency reparandum items, or items that are part of low-frequency bigrams are not more or less recognisable as repairable than high-frequency ones. The cloze-based measure of repair word predictability is significantly correlated with several frequency measures, most strongly the repair item's following bigram (r=0.283, p<0.001).

Thus, repair items that are easily guessed from context tend to be part of high-frequency phrases.

Turning to relationships between the semantic classification and probabilistic measures, Fig.1 and Fig.2 illustrate that appropriateness, factual error and linguistic error repairs are associated with distinct ‘predictability profiles’. Fig.1 shows that most appropriateness repairs involve reparanda that are difficult to spot (repair predictability 0), and most of these have resolutions that are difficult to predict (repair item predictability 3). Linguistic error repairs mostly involve errors that are easy to spot (repair predictability 3), and most of these have obvious resolutions (repair item predictability 12). Factual error repairs are of two types: errors are either easily spotted and resolved, or neither.

Figure 1: Area histograms for repair item predictability split by repair predictability (horizontal) and repair type (vertical).



The pattern in Fig.1 is arguably not surprising. For factual error repairs, relevant factual information can be present in prior discourse (high contextual predictability), or absent (low). Linguistic errors involve ill-formedness that is mostly easily recognised and resolved whatever the context. Appropriateness repairs involve subtler rephrasings: here, reparanda are not clearly erroneous, and it is often far from obvious to the listener what the speaker might consider a more appropriate phrasing.

Figure 2: Mean frequency values for the repair item by repair type.

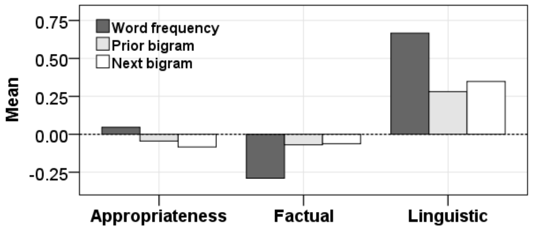


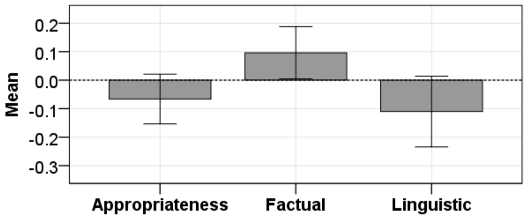
Fig.2 shows that linguistic error repairs are associated with substantially higher lexical frequency means than factual error repairs and appropriateness repairs. The relationship between the latter varies depending on the frequency measure. Again, this pattern is not surprising: at least half of the linguistic error repairs involve the erroneous selection of high-frequency grammatical words, such as prepositions and verbal particles.

3.2. Modelling prosodic marking

I assessed the value of the semantic and probabilistic variables for modelling the prosodic marking judgements through linear mixed effects modelling in R (*lme4* package). I constructed a base model with speaker identity as a random effect, and assessed through log-likelihood comparison whether the addition of any candidate predictors resulted in significant improvement of model fit.

The analysis confirmed that a semantic classification of repairs following [5, 6, 10] is a significant predictor of prosodic marking (improved fit over base model: χ^2=9.717, df=2, p=0.008). By contrast, none of the probabilistic variables showed any predictive value, whether added to the base model, as an interaction term with the semantic classification, or—following residualisation where relevant—as a second main effect. Therefore, the final model contains only a random effect for speaker (sd=0.157) and a fixed effect for repair type (df=2, F=4.923). The effect of repair type is visualised in Fig.3, in which the frequency of prosodic marking is represented by the residuals of the base model. Fig.3 shows that in line with [6], factual error repairs are more often prosodically marked than appropriateness repairs. However, linguistic error repairs are least often marked. The difference between appropriateness and linguistic error repairs is not significant (Tukey’s HSD: p=0.858); the differences between both and factual error repairs are (p=0.030, p=0.031 respectively).

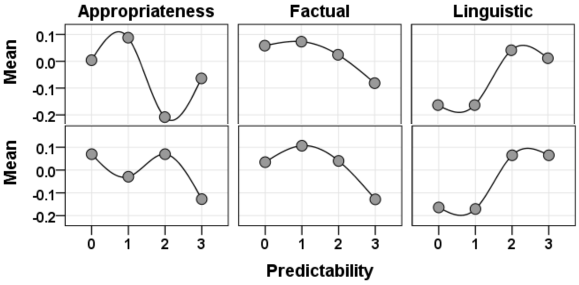
Figure 3: Mean frequency of prosodic marking by repair type. Error bars represent 95% confidence intervals.



The pattern in Fig.3 rules out a straightforward relationship between prosodic marking and predictability: appropriateness repairs and linguistic error repairs are at opposite ends of the probabilistic spectrum (see Fig.1), but are equally likely to be produced with marking. Splitting the data set according to semantic repair type does not reveal further significant patterns: notably, when factual error repairs—which span the entire probabilistic spectrum (see Fig.1)—are considered alone, again none of the probabilistic variables emerge as significant predictors of prosodic marking.

Nevertheless, the cloze-based measures of repair and repair item predictability do reveal interesting trends. These are illustrated in Fig.4, in which the frequency of prosodic marking is represented by the residuals of the final model above, and repair item predictability values are binned to fit a four-point scale. Fig.4 shows that for appropriateness and factual error repairs, repairs associated with low predictability are on average more frequently marked than high-predictability ones. For linguistic error repairs the reverse is true: the linguistic errors that are most easily identified and resolved are most consistently produced with marking.

Figure 4: Mean frequency of prosodic marking by repair predictability (top) and repair item predictability (bottom), split by repair type.



4. DISCUSSION

This study assessed the relevance of predictability for the distribution of prosodic marking in lexical self-repair. A crucial question was whether probabilistic measures add to, interact with or outperform a semantic classification of repairs in modelling marking judgements. None was the case.

The effect of the semantic classification seems at odds with [6]: appropriateness repairs are less frequently marked than factual error repairs, but *more* than linguistic ones. Conflating factual and linguistic error repairs, as in [6], would result in non-significance for the appropriateness–error factor, as in [10]. However, the pattern can be understood in semantic and probabilistic terms. While factual error repairs involve more semantic contrast than appropriateness repairs, linguistic error repairs arguably involve none: the speaker simply gets the grammatical construction of a phrase wrong the first time. As such, these repairs are akin to phonological repairs, which are rarely marked [2, 5]. The lexical items involved also tend to be highly frequent and predictable in context—not highly informative.

The observed difference between appropriateness and factual error repairs provides support for the idea that speakers use prosodic marking to highlight salient information [1, 5]. While appropriateness repairs are associated with low predictability, the effect cannot be reduced to a probabilistic one: among these repairs, probabilistic measures remain non-significant predictors of marking, although cloze-based measures reveal trends in the direction consistent with [5]. It seems plausible that factors not considered in this study, such as discourse-functional ones, further constrain speakers' choices.

Interestingly, the trend for linguistic error repairs is consistent with the idea that marking is a response to conspicuous errors [3]: here we find a positive correlation between predictability and the likelihood of marking. This suggests that speakers' motivations for prosodic marking in repair depend on whether semantic contrast is involved (marking to highlight correct information) or not (marking to divert attention from error). A question for further research is what trend is observed for phonological repairs.

5. ACKNOWLEDGEMENTS

Part of this research was supported by the ESRC grant RES-061-25-0417.

6. REFERENCES

[1] Calhoun, S. 2010. How does informativeness affect prominence? *Lang. Cog. Proc.* 25, 1099–1140.

[2] Cutler, A. 1983. Speakers' conceptions of the function of prosody. In Cutler, A., Ladd, D.R. (eds), *Prosody: Models and measurements*. Berlin: Springer, 79–91.

[3] Goffman, E. 1981. *Forms of talk*. Oxford: Blackwell.

[4] Kapatsinski, V. 2010. Frequency of use leads to automaticity of production. *Lang. Speech* 53: 71–105.

[5] Levelt, W.J.M. 1989. *Speaking: From intention to articulation*. Cambridge, Mass.: The MIT Press.

[6] Levelt, W.J.M., Cutler, A. 1983. Prosodic marking in speech repair. *J. Semantics* 2, 205–217.

[7] Miellet, S., Sparrow, L., Sereno, S.C. 2007. Word frequency and predictability effects in reading French. *Psych. Bull. Review* 14, 762–769.

[8] Nakatani, C.H., Hirschberg, J. 1994. A corpus-based study of repair cues in spontaneous speech. *J. Acoust. Soc. America* 95, 1603–1616.

[9] Pan, S., McKeown, K., Hirschberg, J. 2002. Exploring features from natural language generation for prosody modeling. *Computer Speech Lang*. 16, 457–490.

[10] Plug, L., Carter, P. 2013. Prosodic marking, pitch and intensity in spontaneous lexical self-repair in Dutch. *Phonetica* 70, 155–181.

[11] Pluymaekers, M., Ernestus, M., Baayen, R.H. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62, 146–159.

[12] Seyfarth, S. 2014. Word informativity influences acoustic duration. *Cognition* 133, 140–155.