Modelling Expectation in the Self-Repair Processing of Annotat-, um, Listeners

Julian Hough and Matthew Purver
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
{j.hough,m.purver}@qmul.ac.uk

Abstract

This paper describes a statistical corpus study of self-repairs in the disfluency-annotated Switchboard corpus which examines the time-linear nature of self-repair processing for annotators and listeners in dialogue. The study suggests a strictly local detection and processing mechanism for self-repairs is sufficient, an advantage currently not used effectively under the bonnet of state-of-the-art automatic disfluency processing. We then show how simple local fluency measures using modified language models can be strongly indicative of repair onset detection, and how simple information theoretic measures could characterize different classes of repairs.

1 Introduction

Statistical language modelling for self-repair has enjoyed good results for accurately detecting edited words within repairs (Heeman and Allen, 1999; Charniak and Johnson, 2001; Johnson and Charniak, 2004; Georgila, 2009; Zwarts et al., 2010; Qian and Liu, 2013). However, these successful systems ignore the classification of the repair's function and interpretation; furthermore the models used are generally computationally complex, over-predictive, and unrepresentative of a listener's incremental interpretation process, raising questions of psychological plausibility.

Beginning with classification, we consider the structure and taxonomy of first-position self-repairs, following the annotation scheme first proposed by Shriberg (1994) and the Switchboard disfluency corpus (Meteer and Taylor, 1995) annotation protocol:

$$\underbrace{\text{John and Bill}}_{\text{original utterance}}\ \underbrace{[\ \text{like}}_{\text{reparandum}}\ +\ \underbrace{\{\text{uh}\}}_{\text{interregnum}}\ \underbrace{\text{love }]}_{\text{repair}}\ \underbrace{\text{Mary}}_{\text{continuation}}$$

In addition to this structural vocabulary, from here on we consider the *repair onset* to be the first word after the (possibly null) interregnum, and the *interruption point* as the transition labelled '+ ' between the reparandum and the repair. Within this schema it is possible to distinguish three main classes of repair:

(1) "But one of [the, + the] two things that I'm really. . ."
Repeat (sw4356)[1]

(2) "Our situation is just [a little bit, + kind of the opposite] of that"
Substitution (sw4103)

(3) ". . . the bank was suing them [for, + { uh,}] because they went to get. . ."
Delete (sw4356)

Intuitively, a repair seems likely to be interpreted as a delete (3) if the following word (the repair onset) has no substitutional relation with its reparandum before the interruption point, having an overriding or cancelling effect; substitutions (2), in contrast, do exhibit some substitutive property or parallelism; and verbatim repeats (1), almost trivially, exhibit complete parallelism.

The interpretation of self-repairs by both annotators assigning bracketing on transcripts, and listeners assigning an interpretation function during dialogue, is not trivial. One could argue that simply checking for verbatim repetition for repeats, syntactic constituent identity for substitutions – see Levelt (1983) – and otherwise positing a delete, is a sufficient classification protocol. However there are many different possible subclasses, and gradient effects are exhibited in judgements of the classification of delete or substitution. As we have found in preliminary annotation experiments, annotators do not agree on each decision. The following examples show possible alternative interpretations (italicized) to the Switchboard annotations:

[1] sw* are conversation numbers in Switchboard.

(4) "and [there's, + *?]* it's] completely generic."
Substitution or delete? (sw4619)

(5) "a matter where priorities are [at, +] placed.*?]*"
Delete or substitution? (sw4360)

In terms of the incremental dialogue semantics of these different forms, as Ginzburg et al. (2013) discuss, there is a broad difference between *forward-looking* (verbatim repeats, filled pauses/editing terms) and *backward-looking* interpretations (reformulation such as (2)). Deletes, utterance-initial forms of which are often called *restarts*, are more destructive than substitutions as they are driven by continuing the original utterance, rather than replacing or modifying the reparandum. A speaker or annotator may infer a more dramatic change of content in processing a delete. Also in on-line detection, as they do not adhere to well-formedness rules (Levelt, 1983), one must use other mechanisms to process them.

We maintain the classification distinction between substitutions and deletes, but the need for gradient judgements between these classes is clear due to the possible different interpretations of (4)-(5). Some repairs are more prototypical of their class than others. We address this in section 5.

The second issue we wish to address is the time-linear way in which people process repairs, a constraint which rule-based disfluency detection models do not prioritise – even if they are embedded in incremental systems – as we will discuss below. In consideration of working memory constraints, it is much more likely that repair operations begin once the repair onset is detected, rather than constantly predicting a reparandum before any disfluency has been encountered. Resolution can still be as fast and automatic as psycholinguistic evidence suggests (Brennan and Schober, 2001), but without maintaining all possible repair paths.

We investigate the intuitions of a local self-repair detection and resolution mechanism with gradient interpretation through a corpus study and language modelling. Our corpus study in section 4 observes the frequency of the three main classes and their subtypes in Switchboard, and the interactions of repair class distribution with features of their local utterance contexts. We then present a potential model of incremental repair detection and and interpretation in section 5, based on information theoretic measures.

2 Previous Work

Corpus analysis of Switchboard Shriberg and colleagues (1994; 1996; 1998) have done extensive work annotating and analysing the Switchboard corpus for repairs, editing terms and filled pauses, using a reliable disfluency annotation scheme (Meteer and Taylor, 1995) (see above). Shriberg (1994) creates a taxonomy of disfluency types: filled pause (FP), articulation disfluency (ART), substitution (SUB), insertion (INS), deletion (DEL), repetition (REP), hybrid disfluency (HYB) and conjunction (CON), the last of which occurs between speaker utterances. HYB is an important member of Shriberg's taxonomy due to the plethora of combinations of these repair operations in Switchboard, as we will show below.

Shriberg (1996) compares the distribution of disfluency types across three different dialogue domains including Switchboard. The most common type in all three domains is FP followed by REP. DEL, defined as a repair containing at least one deleted word with no insertions or substitutions, and SUB, defined as a repair having at least one substitutive relation to the reparandum with no deletes or insertions, were ranked 3rd and 4th in Switchboard respectively.

In terms of incremental processing, Shriberg showed an interaction between the position of the interruption point and the disfluency type: per-word rates by position showed that the three most common disfluencies (FP, REP, and DEL) were much more likely to occur in initial position than in medial position. The remaining types appear to be roughly equally likely in initial and medial positions. Furthermore Shriberg and Stolcke (1998) investigate retraces, which are either verbatim repeats or repairs with one or more repeated words. Fitting parameters over the entire disfluency-tagged corpus, there is a logarithmic decay in the likelihood of retracing back one more word as the number of words since the last utterance or repair boundary increases. Speakers rarely retrace more than one or two words. This relationship supports a claim for a very local strategy for repair resolution.

Statistical self-repair detection In state-of-the-art self-repair detection on transcripts, Qian and Liu (2013) achieve the best reported performance on the Switchboard disfluency test corpus, achieving an f-score for detecting reparan-

dum words of 0.841. They use a three step detection system using weighted Max-Margin Markov (M^3) networks: (1) detection of edit-terms/fillers/interregna (2) detection of reparandum words, and (3) refining the previous steps, using a cost-sensitive error function. Georgila (2009) introduces a post-processing method of Integer Linear Programming (ILP) to improve overall accuracy of various off-the-shelf methods, reporting an f-score for detecting reparandum onset words at 0.808 and repair onsets at 0.825 for a CRF model. While these results are impressive, the systems do not operate incrementally: they maximise the overall likelihood of tag sequences in utterances, using utterance-global constraints, rather than focussing on incremental accuracy.

Zwarts et al. (2010) describe an incremental version of Johnson and Charniak (2004)'s noisy channel model. The detector uses a bigram language model trained on roughly 100K utterances of reparandum-excised Switchboard data for its "cleaned" language model. Its channel model is a statistically-trained S-TAG which has simple reparandum-repair alignment rules for its non-terminals (copy,delete,insert,substitute), parsing all possible repair structures for a given utterance hypothesised in a chart, before pruning the unlikely ones. It performs equally well as the non-incremental model by the end of each utterance, achieving an f-score of 0.778 for the Switchboard disfluency task, and is modified to make detections early. They report the novel incremental evaluation method of *time-to-detection* for correctly identified repairs, achieving an average of 7.5 words from the start of the reparandum and 4.6 from the start of the repair phase, longer than the average repair length. They also introduce *delayed accuracy*, a word-by-word recall evaluation of the gold-standard disfluency tags from the point reached utterance so far, reporting recall in one word histories being 0.578, steadily increasing word-by-word until 6 words back where it reaches 0.770.

An earlier incremental system was Heeman and Allen (1999)'s multi-knowledge source approach which employs templates of repair structures within a complex incremental language model. This performs slightly worse than the noisy-channel approach above at detecting reparandum words (recall 65.9% and precision 74.3%), with the sparseness of the data providing problems for templates– the fact that the repeat sequence, w_1, w_1 is the most common repair structure may be very useful for an incremental classifier, but there is a long tail in the distribution of repair structures: they report that 1,302 modification repairs (non-deletes) take on 160 different repair structures in the TRAINS corpus, with only 47 (29.4%) occurring at least twice. To combat this they use over-prediction of templates, initially providing high recall with low precision, then filter out unlikely candidate repair structures using lexical, POS and intonation features. They include a feature encoding that a repair has already been detected in the utterance: in TRAINS, 35.6% of repairs overlap. Utterance-initial cancelling repairs (re-starts), were particularly problematic to identify – we suspect through lack of POS- or word-level parallelism and available templates, which can be exploited in repeats and substitutions, but not for deletes. Heeman and Allen also report very high accuracy for detecting discourse markers/editing terms (both as interregna and as forward-looking repairs), identifying 97% of them with 96% precision.

3 Approach: locally triggered repair detection and classification

In the popular automatic detection task, while incremental systems exist, they use over-prediction, large chart storage and filtering (Zwarts et al., 2010; Heeman and Allen, 1999). A parsing chart used solely for disfluency structures positing every possible repair path grows approximately cubically with the length of the utterance. Also, (Zwarts et al., 2010)'s TAG parser also has a run-time complexity of $O(N^5)$. This complexity blow-up seems cognitively implausible, particularly given the relative sparsity of repairs. In addition, these approaches cannot easily deal with processing embedded repairs realistically, as a stack of charts would be required, further increasing complexity– consequently these are ignored in training (Johnson and Charniak, 2004). Rather than positing all possible repair alignments, intuitively, a listener is almost certain an utterance is a non-repair before the repair onset, so a backtracking mechanism employed upon interruption point detection seems more plausible. A more strictly incremental detection should improve responsiveness (time-to-detection) too.

The clear omission in state-of-the-art systems

is repair classification. We assume dialogue participants are sensitive to the function of a repair for several reasons. One may make direct use of semantic dependencies in substitutions such as "I saw [one, + {no,} two] men", but may draw more pragmatic and turn taking inferences about utterance-initial deletes (restarts). Also, recognizing whether your dialogue partner sits either side of the statistically significant divide between "repeaters" and "deleters" (Shriberg, 1996) may help alignment. Classification's obfuscation in the standard NLP disfluency task is perhaps due to its lack of clarity in definition. Verbatim repeats withstanding, as mentioned above there is often discrepancy between human annotations, suggesting gradient effects; finding a system that can reliably classify the extent of the repair and its function incrementally is a difficult challenge.

Given the problems with the various approaches, we are motivated to find a psychologically plausible incremental method for processing speech repair types by considering the time-linear order in which listeners receive the incoming acoustic signal and then react:

1. Detection of the *interruption point*, triggered via some combination of a partial word, an editing term forming an interregnum or characteristics of the repair onset.
2. Estimation of *reparandum start* position through some backward-looking process.
3. Possibly simultaneously with (2), estimation of the *repair end*, via detection of a further repair, a fluent continuation or the end of the utterance; interleaved with (1) and (2), the repair's *classification*.

In the remainder of this paper we present a corpus survey in section 4 and a proposed approach for modelling repair in section 5 investigating these stages.

4 Self-Repair Distributions in Switchboard

Our initial repair distribution study uses the standard Switchboard training corpora (all conversation numbers sw2*,sw3* in the Penn Treebank III release), plus the non-Treebank Switchboard files, giving a total of 972 transcripts, ∼196,600 utterances, ∼1.28M words, from which we extract 40,485 self-repairs based on the annotations.

The base-rate likelihood of a given word beginning a repair onset is p= 0.0366, that is on average once every 27.3 words of speech.[2] We do not distinguish between repairs crossing utterance boundaries and those marked within an utterance unit, treating them both as first-position within one continual stream, however the difference between these two types would be interesting to consider in further study.

Repair taxonomy by alignment To investigate the distribution of the different types of repair, we follow Johnson and Charniak (2004) in their use of minimum string-edit distance alignment. Ignoring a handful of backwards-looking disfluencies which are annotated within editing term sequences, our aligner classifies 40,364 examples. It operates by mapping each reparandum word to a repair word, where each word must receive at least one alignment with the best possible score. In addition to their alignment categories we introduce COMPLETE_PARTIAL, which aligns prefix→complete word relations such as "j- + just". We used the following scores to ensure that 'weaker' substitutional relations are replaced by stronger ones: REPEAT:6, COMPLETE_PARTIAL:5, SUB[same POS]:4, SUB[same POS first letter]:3, SUB[arbitrary]:2, DELETE:1 and INSERT:1. We decided that as COMPLETE_PARTIAL is a partial repeat it should be selected as a stronger alignment over a SUB[same POS].

The most frequent aligned structures extracted are shown in Table 1: we split the structures between the broad classes of verbatim repeats, substitutions and pure deletes (no repair phase annotated), in order to get the most prototypical deletes as judged by the Switchboard annotators.

1139 different alignment sequence types were found, with only 38.9% of types occurring at least twice, a figure higher than Heeman and Allen (1999)'s reported 29.4%, most likely due to a bigger corpus size. As can be seen, the majority of types are within substitutions, which have a long tail of compound types – the 10 example substitutions shown only constitute roughly half of all substitution occurrences. Deletes were the rarest,

[2] We exclude the first word of every utterance that is not a continuation, as you cannot begin a disfluency repair initiation across these boundaries. ∼100 repairs' repair onset occur at the same word as an embedded repair, so simply dividing the number of word transitions by the number of repairs annotated would give a slight, but insignificant, boost in raw likelihood.

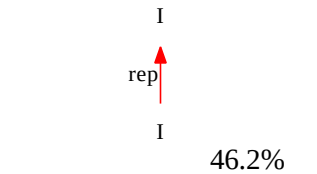
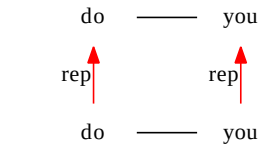
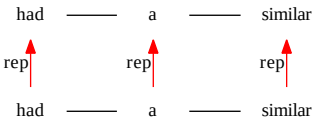
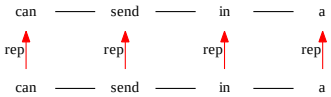
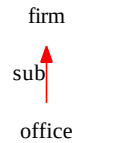
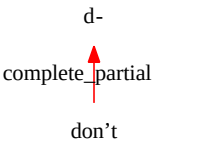
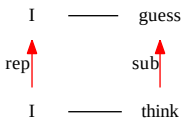
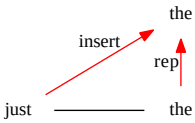
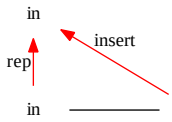
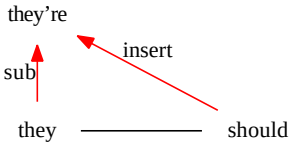
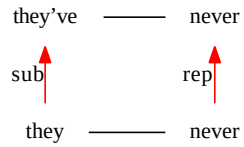
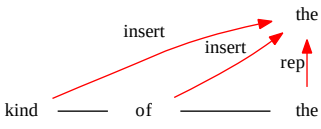
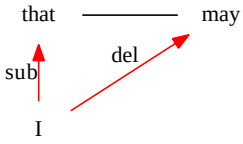
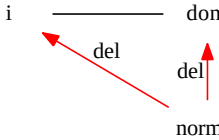
Repair class	Most Frequent repair types (% overall repairs)
Repeats (56.79%) +interregnum=11.96% of class; reparandum=1.23 (std=0.53, power $y = 1.7229x^{-4.425}$, $R^2 = 0.9565$);	 46.2%  8.2 %  1.5 %  0.3%
Substitutions (36.55%) +interregnum=18.65% of class; reparandum=1.78 (std=1.16, power $y = 1.0454x^{-2.593}$, $R^2 = 0.9227$);	 10.2%  3.3%  1.8%  1.4%  1.3%  1.0 %  0.9%  0.9 %  0.7%  0.7%
Deletes (6.66%) +interregnum=0.7% of class; reparandum=1.35 (std=0.88, power $y = 0.938x^{-2.995}, R^2 = 0.9956$);	 5.0%  0.8%

Table 1: Distribution of the most frequent repair disfluencies in Switchboard

conflicting with Shriberg (1996), but mainly due to our definition covering pure deletes only.

While building a rule-based repair grammar is not what we advocate in this paper, it is worth noting the observed alignment sequences can be compressed into 194 different operation sequence pairs such as [*SUB*(r_{m-i}-R_{n-j}) *REP* (r_m-R_n)], in this case representing a substitution alignment from i words back from current reparandum index m to a repair word j words back from current repair index n, followed by a repetition alignment between the current indices. In terms of coverage, due to the sparsity of most alignment sequences, the strength of Johnson and Charniak (2004)'s generative S-TAG grammar approach over a template based one (Heeman and Allen, 1999) becomes clear – for example the approach allows the most frequent repair type, repeats, to have high likelihood within a repair 'grammar', regardless of their length.

Reparandum lengths First-turn repairs tend to be very short, with a mean reparandum length of 1.44 (partial) words (pop. st.dev = 0.88). As with many linguistic phenomena, their length distribution can be characterized as an inverse power law: a function $y = 1.7197x^{-3.61}$, where x is the reparandum length in words and y is the average relative frequency of that length, has a goodness-of-fit R^2 = 0.9635 up to length 9. Reparanda of 1 or 2 words account for 90.8% of repairs and lengths 1-3 account for 96.5%. Repeats (1.23 words) and deletes (1.35 words) are significantly shorter than substitutions (1.78 words), which also exhibit a shallower power-law decay – see Table 1 for the figures.

With the vast majority of reparanda being 1-3 words long, a very local model of context could be used to capture them. As mentioned, previous approaches using sequence-based language models in combination with repair grammars and templates have had success, but there is scope for incorporating repair detection more directly into an n-gram model (though not necessarily through Hidden Event Language Models (HELMs) (Georgila, 2009), which require longer contexts and more training data). Furthermore, as Shriberg and Stolcke (1998) showed, the likelihood of retracing back one more word in retraces decays logarithmically with the number of words into a fluent word sequence, so the need to store all possible reparandum sites before having heard an interruption point seems unnecessarily complex: a locally triggered recovery mechanism does not have far to backtrack. Repeats and deletes are frequently short so their repair onset and reparanda will often fall within a bi- or trigram: for example, presuming perfect interregnum and edit term recognition, a trivial repeat-word feature $w_i = w_{i-1}$ captures 46.2% of all repairs. Use of such local alignments may yield high precision, but we need a more general way of detecting interruption points in a local n-gram context which can also capture longer repairs, as will be discussed below.

Embedded repairs 11.9% of all repairs are embedded inside a longer structure – this divides between 9.9% chaining repairs, embedded within the reparandum phase as in (6), and 2.0% nested within the repair phase of a longer repair.[3] While these appear to need more complex resolution mechanisms, which is presumably why they are ignored in the training phase and evaluation of automatic disfluency systems, they need not be processed as hierarchically embedded structures by listeners on-line. They are frequently short, with mean reparandum 1.28 words long (std=0.67), and so can be resolved very locally, again in a short n-gram context, and may provide an immediate feature for following repair onsets. Intuitively an interruption point indicates speaker trouble, so the likelihood of a consequent interruption point in the following word transitions increases.

(6) " [[This, + it,] + they] are really. "
Embedded chaining substitution- (sw3389)

Partial words as interruption point indicators The most reliable lexical indicator of a repair onset is a preceding partial word. According to the transcripts, the likelihood of a repair onset following a partial word that is not utterance-final is 0.925, boosting the likelihood significantly more than the presence of an interregnum, as will be discussed below. Furthermore, the remaining 0.075 of probability mass for continuations, upon inspection, look like mis-transcriptions. Reparandum-final partial words are present in 10.4% of repairs. Furthermore, the completion of a single partial word is one of the most frequent repair structures (3.3% of all repairs). The probability of the partial word

[3] While Shriberg (1994)'s thesis and Meteer and Taylor (1995)'s annotation attempted to formalise these, they remain a problem for consistency of annotation- it is not always clear whether they should be annotated as nested or chaining.

being a deleted reparandum also rises from the overall average rate 0.066 to 0.171.

This is clearly a very useful feature for detection and classification. Charniak and Johnson (2001) posit an optional phase between the reparandum and the interregnum called the 'free-final', consisting of a sequence of partial words of any length, which, when used as a training feature for an edited words classifier, can improve the detection of repairs. Subsequent work does not use partial words in an attempt to simulate a more realistic testing situation for dialogue systems. While we cannot make direct predictions here without the acoustic data, we investigate how a simple word completion predictor could be a fair approximation to an annotator's incremental processing in section 5.

Interregnum vocabulary Another incremental indication of repair, which has been established in previous empirical work (Clark and Fox Tree, 2002) and in formal models of dialogue (Ginzburg, 2012), is the presence of a conventional editing term for signalling speaker trouble. The editing signals that constitute most repair interregna have a characteristic vocabulary, a fact Heeman and Allen (1999)'s system exploited to detect them with almost perfect accuracy.

In Switchboard, only 13.9% of revision repairs have an interregnum, so it is not a strong repair indicator, which is surprising given its important role in formal and empirical models. However, if one is identified correctly, its presence signals information about the type of upcoming repair: the likelihood of a substitution rises to 0.499, and the likelihood of a delete reduces to <0.01, which could be due to deletion's more destructive semantic 'cancelling' function on the reparandum. There are more substitutions with interregna than repeats in raw frequency and significantly more relative to their class size (2752/14755 (18.65%), versus 2741/22921 (11.96%) $\chi^2_{(1)}$=322.9,p$<$0.0001).

Interregna share a virtually identical vocabulary to editing signals in the more common *abridged* (Heeman and Allen, 1999) or *forward-looking* (Ginzburg, 2012) repairs which comprise an editing signal followed by a fluent continuation to their preceding context, rather than a disfluent one. Focussing here on interregnum vocabulary distributions, we obtain the probabilities in the below table, showing the predictive power of the vocabulary item and its relative frequency within all repairs. The filled pause "uh" and discourse marker "you know" are the most indicative, increasing the probability of a repair from the base rate to 0.155 and 0.1 respectively. These two items are also the most frequently occurring within repairs (9.0% and 2.6% of repairs have them, respectively). The lack of predictive power even the most frequent interregna forms have to predict repair means interregnum presence does not provide a reliable feature for detection on its own; however as it has significant interaction with repair type, it is a useful feature for repair classification.

form	p(repair|form)	p(form|repair)
(fluent word)	0.037	0.861
"uh"	0.155	0.090
"you know"	0.100	0.026
"well"	0.080	0.006
"I mean"	0.074	0.005
"um"	0.061	0.003
"yeah"	0.038	0.002
"or"	0.017	0.002
"like"	0.014	0.003
"so"	0.005	0.001
"actually"	0.025	0.001

5 Language models for on-line repair processing

Having observed some distributional properties of the form of self-repairs that could contribute to on-line detection and classification tasks, we now introduce a simple information theoretic model which incorporates some of them, including local repair detection based on language model probability and partial word presence. This model can be used orthogonally to alignment approaches discussed above, and should provide scope for more efficient, realistic and robust implementations.

We model the task of listeners and annotators as representing the following constituents of a repair, ignoring interregna and other editing terms:

$$...w_o^N[w_{rm}^1...w_{rm}^N + w_{rp}^1...w_{rp}^N]w_c^1... \quad (7)$$

Intuitively and in accordance with the processing order outlined in section 3, the first detection problem is recognizing the repair onset w_{rp}^1 (or w_c^1 for deletes). For this we intuit the most important factor is syntactic disfluency, that is, violation of syntactic expectation. Following a detection of this violation, the task is to find the start of

the reparandum – which can be seen as maximising the fluency of a sequence including $w_o^N w_{rp}^1$ – while simultaneously beginning to compute the repair's parallelism to the reparandum onset w_{rm}^1. The final task is to find the repair end w_{rp}^N (or w_c^1 for deletes) and classify the repair through computing its parallelism to the reparandum up to its end w_{rm}^N. We discuss the tools we use to model violation of expectation and parallelism below.

Fluency measures for incremental repair onset detection We require a language model that can predict which word, or class of words, hearers are likely to hear next in on-going dialogue. Although we currently lack robust large-scale predictive incremental parsers – though see (Eshghi et al., 2013; Demberg et al., 2013) for on-going efforts – we can use an approximation to incremental lexical and syntactic fluency with n-gram language models and insights from recent work on modelling grammaticality judgements (Clark et al., 2013). We train a trigram model with Kneser-Ney smoothing (Kneser and Ney, 1995) as our principal default fluency measurement p^f:[4]

$$p^f(w_i \mid w_{i-2}, w_{i-1}) = p^{KN}(w_i \mid w_{i-2}, w_{i-1}) \quad (8)$$

We can define an additional measure of fluency based on the insights of the frequency of partial words at interruption points in section 4. We train a simple word completion model $p^{complete}(w|w_i)$ which operates on any annotated partial word prefix w_i to provide a distribution over possible completions, and thus the most likely completion (based on the prefix and unigram co-occurrence). For detection purposes, we make the realistic assumption that w_i can only be interpreted as an abandoned partial word after having encountered the following word w_{i+1}, which as the corpus study suggested is almost certain to be a repair onset w_{rp}^1. As opposed to leaving the partial word as unknown vocabulary we can instead define a probability distribution of the completion probability of each word in the vocabulary. So for a partial word w_i, the likelihood of w being its corresponding complete word at the time of interruption is:

$$p^{fluent}(w \mid w_{i-2}, w_{i-1}, w_i) = \frac{1}{Z} \times p^{KN}(w \mid w_{i-2}, w_{i-1}) \times p^{complete}(w \mid w_i) \quad (9)$$

where Z is a standard normalisation constant to ensure that $\sum_{w \in Vocab} p^{fluent}(w \mid w_{i-2}, w_{i-1}, w_i) = 1$. The probability $p^{\hat{fluent}}$ of most likely completion of w_i is then:

$$p^{\hat{fluent}} = \max_w p^{fluent}(w \mid w_{i-2}, w_{i-1}, w_i) \quad (10)$$

The intuition here is that when they encounter a partial word hearers attempt to find the most likely fluent word that both maximises its likelihood to be its completion and also of being a continuation of the two preceding words. If we encounter "yes I remem-", the probability of the completer's best guess will not be as low as if it was unpredictable, such as after an utterance initial "T-". When w_i is partial we use $p^{\hat{fluent}}$ in (10) for our fluency measure p^f, otherwise defaulting to our normal p^{KN} model.

Syntactic fluency measures Use of a standard n-gram model conflates syntactic with lexical predictability. To remove lexical effects and focus on syntactic effects only, we normalise for lexical probabilities by following Clark et al. (2013)'s use of Weighted Mean Logprob (WML). WML divides the logprob of the raw probabilities of all the trigrams in the utterance so far over the summed logprob of the component unigrams, normalising by the length of the utterance so far. We intend to use this incrementally and within local trigram windows rather than for full utterances. So at word w_i, we define our syntactic fluency measure as:

$$WML(w_{i-2} \dots w_i) = \frac{1}{n} \frac{logp^f_{TRIGRAM}(\langle w_{i-2} \dots w_i \rangle)}{logp^f_{UNIGRAM}(\langle w_{i-2} \dots w_i \rangle)} \quad (11)$$

Repair classification by entropy measurement If a low WML measure or low p^f can indicate disfluency, a listener or annotator would then want to compute how similar two contexts were in order to infer the class of repair. To do this using trigram contexts we need a distribution of continuations after each word in repair utterances to be available, which we will refer to as $\theta^f(w \mid w_{i-1}, w_i)$. We can then take the entropy $H(\theta^f)$ to give us a measure of uncertainty in the distribution.

To measure syntactic and lexical parallelism between two words we measure the Kullback-Leibler (KL) divergence (relative entropy) between two different distributions of θ^f. This measure of parallelism will be particularly useful for

[4] Many thanks for use of the excellent code attached to Clark et al's paper, available for download at `http://www.dcs.kcl.ac.uk/staff/lappin/smog/`

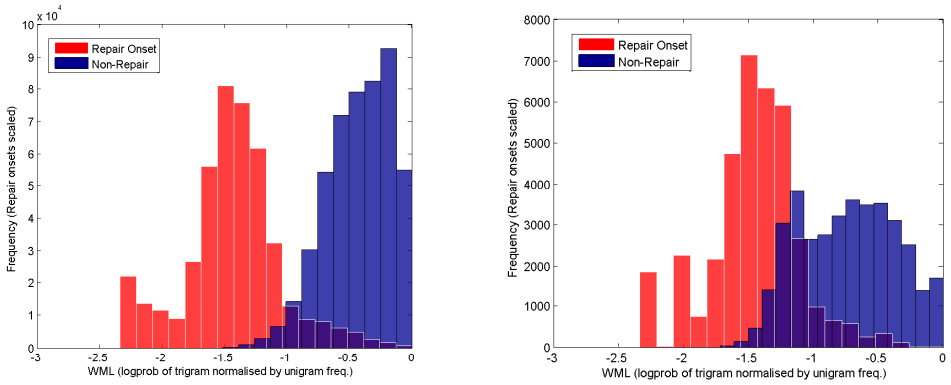


Figure 1: WML fluency measure for training data (left) and heldout data (right)

classification when comparing the θ^f of reparandum and repair boundary words, as will be explained below.

Hypotheses For the incremental processing of self-repair detection and classification, in terms of our fluency and parallelism measures, we hypothesise the following:

1. **Detection**: Repair onsets w^1_{rp} with their context will have significantly lower mean p^f values than non-repair transition trigrams (lower lexical-syntactic probability), and exhibit considerably bigger drops in WML (lower syntactic probability) than other fluent trigrams in the utterance so far, caused by a partial word followed by a fluent one, or other syntactic disfluency.
2. **Reparandum start identification** Processing the utterance with the reparandum removed appropriately will significantly increase the WML of the utterance so far (similar intuition to the noisy channel approach), more so than other hypotheses for w^1_{rp}.
3. **Classification** For repeats, the KL divergence from the continuation distribution after the reparandum's first word, i.e. $\theta^f(w \mid w^N_o, w^1_{rm})$, and that of the repair onset and its cleaned context before the reparandum, i.e. $\theta^f(w \mid w^N_o, w^1_{rp})$, will trivially be 0 in repeats and repeat-initiated substitutions, will be greater for other substitutions and higher still for deletes.
4. **Partial word repair classification** We predict repairs with reparandum-final partial words w^N_{rm} with high entropy over possible completions θ^{fluent} (see equation (9)) will be interpreted as deletes rather than substitutions- in deletes the high uncertainty of predicted complete word is interpreted as 'cancelled'.
5. **Repair end detection/final classification**: In repeats, the continuation distribution at the reparandum-final word w^N_{rm} (i.e. $\theta^f(w \mid w^{N-1}_{rm}, w^N_{rm})$) will be maximally close to that at the repair-final word w^N_{rp} (i.e. $\theta^f(w \mid w^{N-1}_{rp}, w^N_{rp})$) with KL divergence 0. In substitutions, the same KL divergence will be on average higher than in repeats (though for compound type repairs ending in repeats this could still be 0), and the KL divergence for deletes should be even higher.[5] Substitutions as a class may vary significantly within this measure and in the KL divergence in hypothesis (3), however one KL divergence should be sufficiently lower than that of an average delete, and one should be higher than 0 due to them not being verbatim repeats.

Experiments At the time of writing we have investigated hypothesis (1) using the standard division for the Switchboard disfluency detection task for training and held-out data (Charniak and Johnson, 2001, *inter alia*),[6] and for now omitting partial words as per the normal task. After training on a cleaned model (reparandum and edit-terms excised) from the standard Switchboard training data (100K utterances, 650K words), which when

[5] We approximate divergence between $\theta^f(w \mid w^{N-1}_{rm}, w^N_{rm})$ and $\theta^f(w \mid w^N_{rm}, w^1_{rp})$ in deletes, due to the lack of a repair phase; the distribution of continuations after the repair onset (first non-reparandum word) is our best approximation of the repair end.

[6] We reserve the normal test data files for future work.

run over the same training corpus with disfluencies included the model assigns a mean WML of -0.432 (std.=0.262) to non-repair onset trigrams and -1.434 (std.=0.388) to repair onsets. Encouragingly on unseen data, the standard held-out data (PTB III sw4[5-9]*, 6.4K utterances, 49K words.) there is still a significant difference: fluent trigrams had a lower mean, -0.736 (sd=0.359) while repair onsets were similar to their training average at -1.457 (std.=0.359)– see Figure 1. We suspect the sparsity of clean data may have caused this shift, so we would expect to see the effect maintain a healthy gap in testing with a larger language models.

6 Discussion

We have described self-repair processing in terms of probabilistic expectation violation and distributional distance in a fluent language model. We argue this could be a more realistic model than alignment driven self-repair detection posited in state-of-the-art computational models, due to its efficiency and lack of over-prediction. The repair onset detection can be triggered with no latency through using a simple language model. We hope to show conclusively in future work that the many different types of repair distinguished by automatic alignment in our corpus study can be captured by our simple information-theoretic model of incremental fluency estimation and local repair.

Acknowledgements Thanks to the SemDial reviewers for their helpful comments and to Shalom Lappin and Gianluca Giorgolo for their code.

References

S. Brennan and M. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

E. Charniak and M. Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the second meeting of the NAACL on Language technologies*, ACL.

H. H. Clark and J. E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73–111.

A. Clark, G. Giorgolo, and S. Lappin. 2013. Statistical representation of grammaticality judgements: the limits of n-gram models. In *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*. Bulgaria. ACL

V. Demberg, F. Keller, and A. Koller. 2013. Parsing with psycholinguistically motivated tree-adjoining grammar. *to appear in Computational Linguistics, Vol. 39, No. 4*.

A. Eshghi, J. Hough, and M. Purver. 2013. Incremental grammar induction from child-directed dialogue utterances. In *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*. Bulgaria. ACL

K. Georgila. 2009. Using integer linear programming for detecting speech disfluencies. In *Proceedings of Human Language Technologies: NAACL 2009*, pages 109–112. ACL

J. Ginzburg, R. Fernández, and D. Schlangen. 2013. Dysfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*.

J. Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

P. Heeman and J. Allen. 1999. Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571.

M. Johnson and E. Charniak. 2004. A tag-based noisy channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the ACL*, ACL '04.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *ICASSP-95*, volume 1, pages 181–184. IEEE.

W. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

M. Meteer and A. Taylor. 1995. Disfluency annotation stylebook for the switchboard corpus. ms. Technical report, Department of Computer and Information Science, University of Pennsylvania.

X. Qian and Y. Liu. 2013. Disfluency detection using multi-step stacked learning. In *Proceedings of NAACL-HLT*, pages 820–825.

E. Shriberg and A. Stolcke. 1998. How far do speakers back up in repairs? A quantitative model. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2183–2186.

E. Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California, Berkeley.

E. Shriberg. 1996. Disfluencies in switchboard. In *In Proceedings of the ICSLP 96*, volume 96, pages 3–6. Citeseer.

S. Zwarts, M. Johnson, and R. Dale. 2010. Detecting speech repairs incrementally using a noisy channel approach. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10. ACL.