Proceedings of Disfluency in Spontaneous Speech (DiSS: Satellite meeting of ICPhS '99) U.C. Berkeley, 30 July 1999, pp. 35-38.

TOWARD A FORMAL CHARACTERISATION OF DISFLUENCY PROCESSING

Dafydd Gibbon and Shu-Chuan Tseng
Universität Bielefeld, Germany

ABSTRACT

Inherent structural characteristics of speech disfluencies are the prerequisite for the fulfilment of detecting and correcting speech disfluencies in spontaneous speech. However, a considerable number of recent research works on speech disfluencies focus on the surface patterns of speech disfluency editing structure, instead of looking into the relations between editing structure, the syntactic structure and the prosodic structure of speech disfluencies. In this paper we present first results of a new line of research, using feature structures modelled by finite state transducers, on the formal modelling of speech disfluencies in unplanned speech, in relation to all three levels of description.

1 INTRODUCTION

Recent studies of speech disfluencies have focussed mainly on exploring the structural characteristics of speech disfluencies, with the goal of developing psycholinguistic models. However, another line of attack is developing: an engineering need to provide robust human language technology systems, with the ability to cope with disfluencies in speech recognition, either as 'noise' or as functional components of speech, or even perhaps to introduce elements of disfluency into speech synthesis in an effort to simulate more natural and intelligible speech.

Empirical studies have shown that disfluencies are not arbitrary but can be characterised systematically. To describe the internal structure of speech disfluencies in spontaneous speech, most approaches have adopted an 'autonomous template model', without considering the relation of disfluency patterns to the syntactic contexts or the prosody of the elements concerned, for example Levelt [9], Shriberg [13], Heeman & Allen [5] and Bear & al. [2]. Heeman & Allen and Bear & al. were concerned with annotation systems for speech data. They used pattern-based detection of one-word and two-word speech repetitions, insertions and adjacent replacements, also without explicitly using syntactic context.

In this paper we present initial results of research into developing a more explicit declarative dimension to disfluency processing, in the expectation that this will make disfluency processing easier to integrate into a modern processing model. After a discussion of the functionality of disfluency we discuss a family of FST (finite state transducer) models for disfluency and illustrate the application of an FST model to data taken from a German instruction dialogue corpus [3]. The results are documented in detail in [14].

2 DISFLUENCY DETECTION

Psycholinguistic interest in disfluency is partly concerned with disfluency production and perception processes *per se*, and partly with the light that dysfunctionalities can cast here, as in other areas of language performance, on representations and processes of perception and production in general.

Current models of disfluency perception are essentially experimental. Lickley & Bard [11], for example, carried out gating experiments with the aim of finding out what kinds of linguistic cue can help human listeners to detect disfluency. Their results indicate that prosodic cues play a more decisive role in the detection of disfluency than explicit lexical cues.

Linguistic methodology is essentially based on the distributional analysis of corpora, whether small and model-directed, or large. This is the methodology of the present approach, in which computational models of patterns in a corpus of unplanned speech [3] are developed. Distributional linguistic or computational linguistic methods yield results which are in principle neutral with regard to of perception (parsing) and production (generation), though perhaps closer to production. Levelt [9], Tseng [14] and others have shown using large corpora that the majority of speech repairs, especially of complex forms, have a regular internal form, for which a *three event model* can be formulated: Levelt [9] used the categories *reparandum* (stretch of speech to be repaired), *editing term* and *alteration* for the three events, and Tseng [14] used the categories *problem item*, *editing phase* and *corrected item* in a related three event model. Levelt's three event model represents the classical template approach to disfluency structure:

Template: <OrigUtt,EdPhase,Repair>
OrigUtt=<X,reparandum,delay>
EdPhase=editterm
Repair=<retrace,alteration,Y>

The original utterance contains the reparandum, the editing phase consists of editing terms and the repair

contains the alteration, which is the correction of the reparandum.

However, in Tseng's data, the majority of complex speech disfluencies turned out to involve *items* which were *phrases*, and which are thus best characterised as *problem phrase*, *editing phase* and *corrected phrase*, where *phrase* is a unit dominated by a syntactic category (such as NP, VP, PP). This distributional result demonstrated the importance of the linguistic unit *phrase* in the production of speech disfluencies, and the need for explicit phrasal models, in contrast to the 'autonomous template' models used in earlier work which did not take phrasal syntactic structure explicitly into account,

Approaches to disfluency modelling within the engineering context of human language processing, Heeman & Allen [5], Bear & al. [2] and Nakatani & Hirschberg [12], have all used template–based annotation systems to label their data. However, more complex processing models have been used. Hindle [6] built a procedural parser to automatically detect and correct syntactic non-fluencies. Langer [8] set up normalisation rules on the basis of finite state automata to detect and correct syntactic speech repairs. Althoff et al. [1] used a finite state transducer as a word lattice parser in a speech recognition system to correct disfluencies in compound words.

In addition to syntactic disfluency modelling, other linguistic categories have been dealt with. The results on prosody by Lickley & Bard [11], have already been noted; Levelt & Cutler [10] also reported that prosodic marking was present in speech repairs, These results were confirmed, with different methodology, by Tseng [14].

From studies such as these, the conclusion can be drawn that regular patterns for the detection of disfluencies are available, and that these regular patterns may be suitable for use in disfluency detection models for cognitive processing, and in disfluency detection components of human language technology systems.

3 DISFLUENCY PROCESSING

The phase of disfluency detection is logically (not necessarily temporally) followed by the phase of disfluency processing (it is conceivable that disfluency signals may trigger hypotheses about possible repairs before the disfluency has completed its editing and alteration phases, either in the speaker or in the hearer).

3.1 Template models. As already noted, in general, disfluency models have been template-based, i.e. finite structures with slot-filler characteristics, as with the Levelt three event model. A recent integrative template-based approach is developed in Tseng [14], in which complex disfluencies in noun and prepositional phrases are formally described.

But while templates express a form of declarative 'observational adequacy', it is necessary to understand their formal properties in order to be able to suggest plausible processing models. As a first approximation, it may be suggested that disfluency templates are finite structures, and therefore by definition trivially describable by regular grammars (equivalently, finite state automata), and that correction mechanisms may be implemented as finite state transducers (FSTs).

3.2 FST models. Empirical evidence shows that although disfluency sequences can be rather short, they are in principle of arbitrary length, so that a finite template model is not helpful, and more general finite state automata with cyclic structures must be considered. General cyclic models are clearly over–powerful; there are narrow performance constraints on length and consequently additional (perhaps statistical) length constraints must be considered.

A number of empirically validated FST models have indeed been proposed, such as Langer's Disfluency Filter model Langer [8], Tseng's Disfluency Repair model [14], and the Broken Compound model of Althoff & al. (1996) [1]. The latter has been operationally validated by in the form of an implementation as a component of a a speech recognition system.

It can be shown, however, that while standard FST models are adequate for many disfluency types, more complex models are also required, which take prosody and linguistic structure into account (cf. also Lickley & Bard [11] for the detection of disfluency, relying on prosodic cues rather than explicit lexical cues), and which go beyond the classical structures of FSTs. This means that syntactic and prosodic contexts of speech disfluencies influence the production form as well as the production length of speech disfluencies. The results of modelling the distributional data as an FST are shown in Figure 1; the relation between straightforward lexico–syntactic information and 'metalocutionary' editing is coded in style of the transition graphs; for the statistical properties of the FST, see [14]; length heuristics are not considered here.

3.3 Multitape FSTs. Formally, an FST (finite state transducer) can be seen as a finite state automaton (FSA) whose transitions are labelled with elements of a vocabulary of pairs or longer tuples, rather than the atomic elements of garden variety FSAs. A standard FSA is said to accept a *regular language*, while a standard FST is said to accept a *regular relation*. The

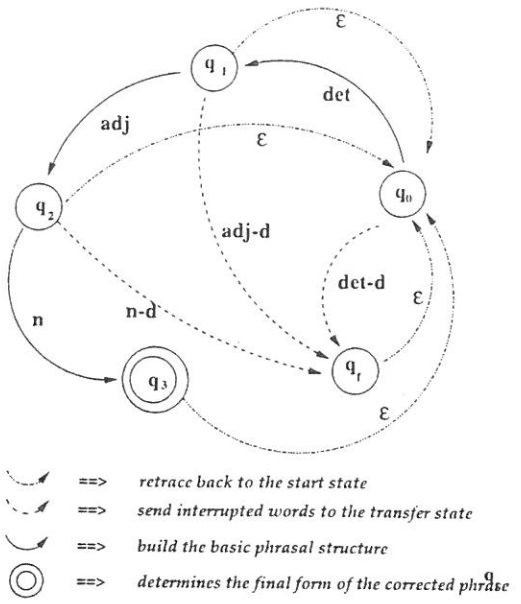

Figure 1: Relation between 'lexico-syntactic' and 'metalocutionary' information.

minimally structured (standard) FST accepts a binary relation, with the left-hand element of each pair in the relation being regarded as an *input* symbol and the right hand pair as an *output* symbol; binarity is not an essential condition, however. FSTs need not be thought of only as input/output devices; they can also be interpreted as processors for parallel streams of information. Kaplan & Kay [7] discuss the application of such FSTs to phonology.

We propose multi-tape FSTs as devices for formalising the relations between the different structural levels involved in the detection and processing of disfluencies, and that the parallel streams of information which are being processed are essentially the following:
lexico-syntactic information stream: reparandum & alteration;
prosodic information stream: pitch & duration;
metalocutionary information stream: editing term & the phonetic (and semantic) *change operations* over reparandum and alteration.

It has been shown by Carson-Berndsen [4] that FSTs can be used to interpret (i.e. represent a model for) an *event logic* model of the sequential and parallel events which make up utterances: constraints between parallel streams are mapped to (underspecified) feature structures, and sequential constraints are mapped to transitions between states of the automaton. This approach has been operationally validated in an experimental spoken language recognition system. The Levelt autonomous template model is a useful abstraction away from details of FST processing, once these details have been established, and the representation of the model as a feature structure can be regarded as a step towards a plausible underlying representation for the third, metalocutionary information stream.

The sequential constraints which map to phonetic sequences in the metalocutionary information stream can be described in terms of concatenation; the treatment of parallel constraints between information streams will be discussed briefly below.

Using a conventional feature structure notation, with co-indexing of structure, to represent the Levelt template as an abstraction from the FST, both phonetic interpretation and semantic interpretation can be related to each other systematically:

$$\left[\begin{array}{ll} \textit{disfluency} & \\ \text{PHON} & = [a] \frown [1] \frown [b] \frown [2] \frown [c] \frown [3] \\ \text{SEM} & = [2]\,([3] \sqcup_{default} [1]) \\ \text{OrigUtt} & = \left[\begin{array}{ll} a & \text{X} \\ 1 & \text{reparandum} \\ b & \text{delay} \end{array}\right] \\ \text{EdPhase} & = \left[\begin{array}{ll} 2 & \text{edit term} \end{array}\right] \\ \text{Repair} & = \left[\begin{array}{ll} c & \text{retrace} \\ 3 & \text{alteration} \end{array}\right] \end{array}\right]$$

The main semantically relevant constituents are marked with numbers, and other elements are marked with letters.

The *syntactic and prosodic control* constraints between parallel information streams can be represented by *association lines*, as shown in Figure 2.

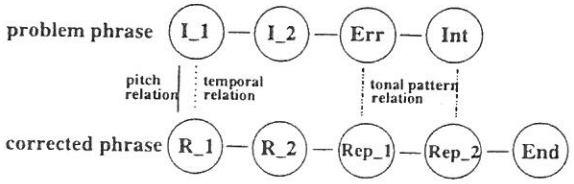

Figure 2: Prosody-metalocutionary synchronisation (association).

But why is semantic intepretation also relevant, and what might be the interpretation of the operator $\sqcup_{default}$ and the functional structure? Just as the phonetic realisation of the reparandum is a fact which remains, and is not in fact — despite current terminology — 'altered', 'repaired', or 'corrected', but simply *supplemented* by the repair, so the semantic

interpretation of the reparandum is also a fact which, particularly in the case of a contradiction (sometimes a 'Freudian slip') or in the case of a co–hyponymous meaning, remains and can be integrated into a complex semantic interpretation of the whole disfluent expression. Examples from the corpus, where semantic interpretation of the reparandum is relevant, are:

`ist bei mir auf der rech--, ich kann es auch umdrehen`

'is on my side on the ri—, I can also turn it round'

The likely semantic interpretation of the reparandum 'rech', 'right hand side' is available to the hearer if needed.

`Uhmm jetzt fügst Du mit dem mit dem an mit den drei mit den fünf Löchern UHMHM mit dem langen Stück`

'er now you start with the with the with the three with the five holes um with the long piece'

The series of abortive repairs yields a set of semantic interpretation hypotheses: is 'the long piece' the same object as the 'five hole' object?

So disfluencies are not just *noise*. The operator $\sqcup_{default}$ is *default unification*, essentially overriding of the meaning of the reparandum by the meaning of the repair (often, but not always, leading to identity) within the lexico–syntactic information stream. The metalocutionary function is qualification of the result of default unification by the the operator from the metalocutionary information stream, for instance by indicating a focus shift.

4 CONCLUSION

On the basis of distributional data collated and formalised in [14], it has been shown that disfluency structures can be represented with formal means which are already in use in computational phonology its applications to the human language technologies.

We suggest that in work in this area, well–understood representation systems and procedures for manipulating them should be used in order to facilitate the integration of 'exotic' facts about speech such as disfluencies into representations of the more familiar parts of the linguistic universe, in particular to syntax and prosody. With a strategy such as this, previous usefully illustrative, but *ad hoc* notations and diagramme styles can be superseded by formalisms rather than notations, which promise both greater generality and precision, and the hope of explanatory power within the context of language representation and processing as a whole.

We believe we have shown the feasibility of such a programme in the present paper; present work marks only a beginning, however, and leaves many gaps, such as symbolic and numerical length constraints, exact mappings to phonetic correlates, or fine details of the semantic interpretation operations, for future research.

REFERENCES

[1] Althoff, F., Drexel, G., Lüngen, H., Pampel, M. and Schillo, C. 1996. The Treatment of Compounds in a Morphological Component for Speech Recognition. In Gibbon, D. (ed.), *Natural Language Processing and Speech Technology.*

[2] Bear, J., Dowding, J. and Shriberg, E. 1992. Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog, In *ACL*, 56-63.

[3] Brindöpke, C., Häger, J., Johanntokrax, M., Pahde, A., Schwalbe, M. and Wrede, B. 1995. Darf ich dich Marvin nennen? Instruktionsdialoge in einem Wizard–of–Oz–Szenario: Szenario–Design und Auswertung. SFB 360 Report 95/16, Universität Bielefeld.

[4] Carson–Berndsen, Julie 1998. *Time Map Phonology: Finite State Methods and Event Logics in Speech Recognition.* Kluwer Academic Press: Dordrecht.

[5] Heeman, P. and Allen, J. 1997. Detecting and Correcting Speech Repairs. In *ACL*, 295-302.

[6] Hindle, D. 1983. Deterministic Parsing of Syntactic Non-fluencies. In *ACL*, 123-128.

[7] Kaplan, Ron M. & Martin Kay 1994. Regular models of phonological rule systems. *Computational Linguistics* 20(3), 331–78.

[8] Langer, H. 1990. Syntactic Normalization of Spontaneous Speech. In *COLING-90*, 180-183.

[9] Levelt, W. 1983. Monitoring and Self-Repair in Speech. *Cognition*, 14: 41-104.

[10] Levelt, W. and Cutler, A. 1983. Prosodic Marking in Speech Repair. *Journal of Semantics*, 2(2): 205-217.

[11] Lickley, R.J. and Bard, E.G. 1992. Processing Disfluent Speech: Recognising Disfluency Before Lexical Access. In *ICSLP*, 1499-1502.

[12] Nakatani, C. and Hirschberg, J. 1993. A Speech-First Model for Repair Detection and Correction. In *ARPA Workshop on Human Language Technology*, 329-334.

[13] Shriberg, E. 1994. Preliminaries to a Theory of Speech Disfluencies. PhD Thesis. University of California at Berkeley.

[14] Tseng, S.-C. 1999. Grammar, Prosody and Speech Disfluencies in Spoken Dialogues. PhD thesis. University of Bielefeld.