Proceedings of ICSLP 2000, Beijing, 16–21 October, 2000.
Paper 01440, Vol. 2, pp. 991–994.
CD-ROM ISBN 7–900618–63–5/G · 54.

Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and Tok Pisin Human–Human ATIS Dialogues

Robert Eklund

Telia Research AB, Farsta, Sweden
NLPLab, Dept. of Computer and Information Science, Linköping University, Sweden

ABSTRACT

This paper studies disfluencies in authentic human–human dialogues in Swedish and Tok Pisin. It is found that while there are no major differences as to types or frequencies on a macro level, there are dissimilarities on a micro level, notably in the characteristics of how prolonged segments are realized. The paper also discusses the results in the light of reported disfluencies in English, German, Ilokano and Tagalog.

1. INTRODUCTION

Current automatic speech recognition (ASR) and human–computer dialogue systems have attained a technological level that allows use in everyday commercial applications, as long as the tasks are sufficiently constrained. In order to allow more open-ended speech input, certain phenomena typical of spontaneous speech need to be modeled. One such phenomenon is the processing of disfluencies, DFs (pauses, truncations, prolongations, repairs, repetitions, etc.). In order to acquire deeper knowledge concerning the role DFs play in human speech production and understanding, they need to be studied across languages, since such studies could point to universal tendencies with regard to DF types, frequencies and distribution. This could in turn point to universal constraints in human speech production, and thus provide predictive power to theories and models dealing with spontaneous speech. In a previous comparative study, Eklund & Shriberg [4] showed that there were many similarities with regard to DF types and distribution between American English and Swedish. Since English and Swedish are quite close, both linguistically and socio-culturally, this paper addresses the universality issue of DFs by looking at two rather more distant languages with little cultural cross-breeding: Swedish and Tok Pisin.

2. TOK PISIN

Tok Pisin is an English-lexicon pidgin/creole language spoken in Papua New Guinea. It is one of the three official languages of this nation counting approximately 800 languages. (The other two official languages are English and Hiri Motu.) Although around 80% of the lexicon are derived from English, the syntax is predominantly Austronesian. Pidgin/creoles are known for simple morphology and Tok Pisin is no exception, but does possess some morphological markers, like the productive *-im* transitive suffix, and the likewise productive *-pela* adjective suffix. For a detailed account, see Verhaar [11].

3. METHOD

3.1. The Corpora

In order to compare similar data from Swedish and Tok Pisin, two corpora of authentic air travel bookings (ATIS) were labeled for disfluencies and analyzed.

Swedish Corpus (SDS) The Swedish corpus was collected during the period February to September 1998 on location at the travel agencies STA Travel and Bennett Travel Bureau in Lund, Sweden, as part of the Swedish Dialogue Systems project. A total of five dialogues/subjects (3 male & 2 female) and 354 utterances were collected (agents excluded).[1]

Tok Pisin Corpus (TP) The Tok Pisin dialogues were collected at Kavieng Airport, New Ireland Province, Papua New Guinea in December 1999 and January 2000 by the author [2]. A total of 39 dialogues/subjects (30 male & 9 female) were collected, counting more than 1,200 utterances (agents excluded). A subset of 654 utterances (3,118 words) have been labeled and analyzed. It comes with the multilingual territory of Papua New Guinea that speakers frequently switch between languages (Tok Pisin, English, local). In this study, only Tok Pisin utterances were included, although English words—given that English is the main lexifier language of Tok Pisin—often appear in otherwise Tok Pisin-only conversation.

3.2. Disfluency Annotation

The corpora were labeled according to an annotation scheme described in Eklund [3]. This system is based on, and consequently similar to, the annotation scheme developed by Shriberg [9]. Both corpora were labeled by the author. In this paper, the following DFs were studied:

Filled pauses (FPs) Also called “filler words” in the literature, most often realized as “eh” or “öh” in Swedish.

Unfilled pauses (UPs) Silent parts in fluent speech. An example would be “I want a … flight to Kavieng.” UPs are often not included in the DF studies. (For a discussion on the ontology of UPs, see Bell et al. [1].)

Prolongations (PRs) Segments which are markedly longer than in normal, fluent speech, e.g. “I want a fffffflight to Madang.”

Explicit Editing Terms (EETs) Words and phrases like “Sorry”, “No, wrong”, and so on.

[1] Merle Horne and Petra Hansson, Lund University, PC.

Truncations (TRs) Interrupted words, either in self-induced repairs or caused by an intervening agent. An example would be "Please book the *fli…*"

Mispronunciations (MPs) Words with the wrong pronunciation, e.g. "I want to make a *veseration*."

Repairs (REPs) A variety of self-corrections, including: substitutions ("I want to find a *train* … *plane* to Malmö."); repetitions ("Please *find me* … *find me* a ticket to Stockholm."); insertions ("I want a ticket … a *cheap* ticket to Port Moresby."), and others. In this paper, each cut-off point counted as one REP, in both simplex and complex (nested) repairs.

4. RESULTS

4.1. Overall DF Rates

The analyzed data are presented in Table 1. Since unfilled pauses (UPs) are often excluded in the literature, but discussed in this paper, the DF counts are given both with and without disfluent sentences containing only unfilled pauses.

Table 1: Summary corpus statistics and DF sentence rates. Figures are given both for all utterances and with one-word utterances excluded. The percentages are given both for all utterances containing DFs and with disfluent utterances only containing UPs excluded.

	SDS	TP
Subjects	5	39
Utterances	354	654
Utts. excl. one-word utts.	224	440
Words	1,854	3,118
Disfl. utts. (total)	61	172
Disfl. utts. (excl. UPs only)	50	113
Disfl. utts. (total) / total utts.	17.2 %	26.3%
Disfl. utts / total utts. excl. one-word utts.	27.2%	39.1%
Disfl. utts. (excl. UPs only) / total utts.	14.1%	17.3%
Disfl. utts. (excl.UPs only) / total utts.excl.one-word utts.	22.3%	25.7%

When all disfluent utterances are included, the difference between SDS and TP is significant, both when the figure is divided with the total number of utterances (p = 0.009, chi-square), and when one-word utterances are excluded (p = 0.033, chi-square). However, if all UP-only disfluent utterances are excluded, there is no statistically significant difference between the two corpora, irrespective of whether one-word utterances are included or excluded. It is hard to draw any conclusions from these figures. The figures are slightly higher than what is normally reported in the literature for human–human conversation, especially in TP, but it must be borne in mind that the recording conditions at Kavieng Airport were less than ideal (cf. [2]). This may have been the case at the two Swedish travel agencies, as well, and given that the two corpora supposedly differ only with regard to language, and pending analyses of more controlled data, we may cautiously assume that there is no significant difference as to overall DF between Swedish and Tok Pisin.

4.2. Specific DF Rates

The next question is whether differences can be found at a more fine-grained level of analysis. Figures for the DFs brokeen down by type are given in Table 2.

Table 2: Summary of DF rates, broken down by DF type. For both corpora, both absolute numbers and percentages are given. Summarized figures are given both for the total number of DFs including UPs and excluding DFs.

	SDS	TP
Total FPs	17	76
FPs / words	0.9%	2.4%
Total UPs	58	264
UPs / words	3.1%	8.4%
Total PRs	8	34
PRs / words	0.4%	1.1%
Total TRs	12	43
TRs / words	0.6%	1.4%
Total MPs	1	5
MPs / words	0.05%	0.16%
Total EETs	1	2
EETs / words	0.05%	0.06%
Total REPs	41	63
REPs / words	2.2%	2.0%
Σ DFs incl. UPs	138	487
DFs incl. UPs / words	7.4%	15.6%
Σ DFs excl. UPs	80	223
DFs excl. UPs / swords	4.3%	7.1%

While there were no convincing differences between SDS and TP on the utterance level, there appear to be differences on the DF-per-word level. Since these differences vary according to the specific DF type, the results will be presented separately. Given the small sample in SDS, it is questionable whether any far-reaching conclusions can be drawn from the observations.

FPs The corpora differ significantly ($p < 0.001$, chi-square). One noticeable difference is the relative proportion of utterance-initial FPs: In TP more the 43% of the FPs begin an utterance, while only 17% of the FPs are utterance-initial in SDS. While it has been shown that FPs might occur word-internally in compounds in Swedish [4] and German [5], and between affixes and roots in Tagalog [7], no such examples of word-internal FPs were found in either SDS or TP.

UPs Once again, the corpora differ significantly ($p < 0.001$, chi-square). UPs are commonly encountered inside words, both in compounds in languages such as Swedish [4], German [5] and Tagalog [7], and inside lexical roots in Swedish [1]. One word-internal UP is found in TP in the compound word "wok … de" (workday). However, no example is found where a UP occurs within a lexical root. In all cases of mid-root UPs, the word is re-started after the silent interval, e.g., "ti … tiket" (ticket), "tr- … traim" (try), "w- … wetlist" (wait list). Although the data are limited, this could point to a constraint, or at least a preference, among Tok Pisin speakers to restart an interrupted word, which could be parallel to the lack of observed word-internal FPs in English [4].

PRs SDS and TP differ weakly (p = 0.015, chi-square). What is more interesting, however, is to take a detailed look on what type of segments are prolonged, and in what position of the word PRs occur. The results are shown in Table 3 and Table 4.

Table 3: PR position in words.

	SDS	TP
% Initial phone	37.5	15.8
% Medial phone	12.5	0.0
% Final phone	50.0	84.2

Although SDS only counts eight PRs, these confirm the observations made in Eklund [3] and Eklund & Shriberg [4], i.e., the roughly 30–20–50 proportions for initial, medial and final position, respectively, found in Swedish and English. Here, Tok Pisin deviates with a roughly 15–0–85 ratio. Similar to the lack of word-internal UPs, there are no instances of word-internal PRs. This could lend further support to the hypothesis that there might be a general unwillingness among speaker of Tok Pisin to employ vocalized hesitation within lexical roots.

Table 4: Phone type of PRs.

	SDS	TP
% Vowel	37.5	44.1
% Consonant +sonorant	37.5	41.2
% Consonant –sonorant	25.0	14.7

The observations made by Eklund [3] and Eklund & Shriberg [4] that all kinds of segments are prolonged in Swedish and English are corroborated in the SDS corpus. As was the case with word position, TP deviates from SDS with regard to proportions, with a stronger preference for vowels and sonorant consonants. A detailed analysis further reveals that while all kinds of segments, even voiceless stops, can undergo prolongation in Swedish, the only non-sonorants that are prolonged in Tok Pisin are the continuants /s/ and /f/, found in the words “yesss” (yes) and “ffffe” (fare). Although the lack of observed prolonged stops in TP does not prove their non-existence in Tok Pisin, this still could hint at phonological constraints.

MPs Mispronounced words are very rare in both corpora, and there is no significant difference between SDS and TP.

EETs Like MPs, explicit editing terms are very rare indeed, and SDS and TP do not differ in this respect.

TRs Truncated words occur in both corpora, but are mainly due to interruptions from the interlocutor, rather than being self-induced. The difference is weakly significant (p = 0.018, chi-square), but no conclusions should be drawn before an analysis that includes the travel agents has been carried out.

REPs There is no significant difference between SDS and TP. Given the limited space of this paper, a detailed account of all subclasses of repairs cannot be given. However, a couple of observations will be made. The maximum number of verbatim repeated words, i.e., the retrace length *n*, is *n*=3 in both SDS and TPs. (An example from TP is “i no *i no* bikpela samting” (It’s not that important), where *n*=2). As can be seen in Figure 1, the lines follow each other closely, and are even more striking if compared with Eklund & Shriberg [4] (where *n*=6).

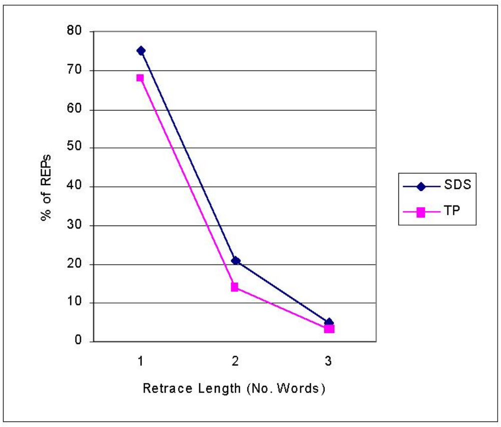

Figure 1: Retrace length distribution in SDS and TP.

TP exhibits a higher number of substituted words in REPs than does SDS, eighteen in TP vs. two in SDS. One particularly interesting example from TP will be shown, where there is a language-switch in the repair: “Tiket *long* … *from* Rabaul to Lae” (Ticket from … from Rabaul to Lae). Here the very general, “catch-all”, preposition “long”, which can mean (aamong other things) both “to” and “from”, is substituted with the English preposition “from”. Such code mixing in repairs are probably to be expected in multi-lingual societies, especially among people with higher education.

4.3. Other Factors

There are of course a plethora of factors that influence the occurrence of DFs. Although this paper cannot study all of them in detail, a brief mention will be made of some parameters that are sure to play a role.

Individual Differences There are huge individual differences, which have not been studied in detail. There is a 1-to-10 difference between the least and the most disfluent subject in SDS if UPs are included, and the difference is 1-to-15 if UPs are excluded. The figures for TP lie within the same range. This is slightly higher than is mentioned by e.g. Oviatt [6] and Bell et al. [1], and a more detailed analysis would be required to explore these differences.

Durational DFs As was observed in Eklund [3], mean duration values of UPs exceed those of FPs, that in turn, exceed those of PRs.

Sentence Length DFs frequency is largely a function of sentence length (cf. [9] [6] [1]). The mean sentence length in SDS is 4.94 words, and the median is 3 words. In TP the figures are 5.17 and 3, respectively. There are also a larger number of very long sentences (above 25 words) in TP.

Agent–Client Interaction As is observed by e.g. Oviatt [6] and Bell et al. [1], DF frequency depends heavily on what type of speech act is carried out. Before more conclusions can be drawn, the interaction and varying roles of the travel agents and the customers must be analyzed.

5. DISCUSSION

There are a number of different phenomena that can be studied in a comparative study such as this, including (but not limited to): (1) DF types; (2) DF type characteristics; (3) DF frequencies; and (4) DF distribution. Moreover, these phenomena must be studied in the light of language taxonomy, syntax, morphology, discourse interaction and socio-cultural influence, among other factors. It goes without saying that a study all of these phenomena requires more space and data than this study is alloted.

As was shown in section 4, there are no great differences between Swedish and Tok Pisin. Overall rates are more or less the same in the two corpora. The clearest dissimilarity seen in the data is the different distribution of word-position and segment type of PRs. One possible explanation for these differences could be the underlying morphotactic constraints of the two languages. Whereas Swedish allows C^3VC^8 syllables (at least in theory), syllable structure in Tok Pisin allows only C^2VC^1 syllables, and even such initial clusters are often split in two by the insertion of epenthetic vowels. Moreover, the phoneme inventory in Tok Pisin is smaller than in Swedish. One hypothesis could be that the more permissive morphology of Swedish has an impact on the realization of prolongations, both with regard to which segments may be prolonged, and also in what word positions they may occur. Naturally, more languages need to be studied before anything conclusive could be said.

PRs are not often explicitly discussed in the literature, which is puzzling, since prolonged segments pose a problem for speech recognizers. Prolonged segments as a linguistic means for hesitation, however, is mentioned within the field of linguistic typology. Streeck [10], discusses “stretched-out sounds” (PRs) in Ilokano—an agglutinating Philippine language—in detail. He also states that “sound stretches […] are thus comparable to the familiar items *uh*, *uhm*, etc. and their cross-cultural variants. Such fillers, however, are almost totally absent from Ilokano conversations. […] Perhaps this is all there is to it: where speakers of English produce a ‘filler’, Ilokano speakers simply stretch out the last vowel before the trouble source. The effect is the same.” (Streeck [10], p. 195.) That Ilokano should lack filler words (or almost so), is challenged by Rubino [8], who presents DF data that include both UPs and FPs. However, Streeck and Rubino agree with regard to the analysis of the important role played by PRs in Ilokano. They also agree on the phonological constraints PRs are subject to, and that they almost exclusively appear on word- or prefix-final vowels.

The positional distribution of the durational DFs (FPs, UPs and PRs) differs between the two corpora. While UPs and PRs are found inside roots in several languages, no such FP example is mentioned in the literature, to the best of our knowledge. On the other hand, FPs are more similar to PRs in their function as acoustically voiced “floor-holders”, in a way that UPs are not. Or, as Streeck puts it: “Ilokano speakers not only continue to vocalize, but also to *speak*: they never cease to say words.” (ibid. loc. cit.) This could imply that PRs are even stronger floor-holders than FPs. In any case, in the light of the studies mentioned above, all the durational DFs should be included in disfluency studies, and their respective roles and distribution be considered, both inter- and intra-linguistically. That the occurrence, distribution and assumed function of DFs vary is clear from the works cited above.

A final point that needs to be made is that this study has focused on observations at the micro level. Further conclusions must take into consideration the full conversational context, which must wait for the full dialogues, including the travel agents, to be analyzed.

6. ACKNOWLEDGEMENTS

I would like to thank the Air Niugini travel agents at Kavieng Airport, Loris Levi, Nianne Kelep and Liza Gabriel. Thanks to Merle Horne and Petra Hansson for providing the SDS corpus. Thanks to Eva Lindström and Carl Rubino for elucidating discussions and comments. Thanks to Jürgen Streeck and Sara Johansson for helping me out in tracking down literature. Finally, thanks to Beth Ann Hockey, Lars Ahrenberg and Eva Lindström for terrific evide… eh I mean *proof* reading!

7. REFERENCES

1. Bell, L. Eklund, R. & Gustafson, J. 2000. A Comparison of Disfluency Distribution in a Unimodal and a Multimodal Speech Interface. *Proc. ICLSP’00*. [These proceedings.]
2. Eklund, R. 2000. Wanpela deitabeis long Tok Pisin bilong baim tiket bilong balus. (An ATIS database in Tok Pisin.) Methodological observations with regard to the collection of authentic human–human data. *Proc. Fonetik 2000*, Skövde, Sweden, 24–26 May, pp. 49–52.
3. Eklund, R. 1999. A Comparative Study of Disfluencies in Four Swedish Travel Dialogue Corpora. *Proc. Disfluency in Spontaneous Speech Workshop*, Berkeley, California, 1 July 1999, pp. 3–6.
4. Eklund, R. & Shriberg, E. 1998. Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues. *Proc. ICSLP’98*, Sydney, 30 November–5 December, vol. 6, pp. 2631–2634.
5. Lüngen, H., Pampel, M., Drexel, G., Gibbon, D., Althoff, F., & Schillo, C. 1996. Morphology and Speech Technology. *Proc. ACL–SIGPHON Conference*, Santa Cruz, pp. 25–30.
6. Oviatt, S. L. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech and Language*, *9*, 19–35.
7. Rubino, C. 1998. The morphological realization and production of a nonprototypical morpheme: the Tagalog derivational clitic. *Linguistics*, 36:1147–1166.
8. Rubino, C. 1996. Morphological Integrity in Ilocano: A Corpus-Based Study of the Production of Polymorphemic Words in a Polysynthetic Language. *Studies in Language*, 20:633–666.
9. Shriberg, E. 1994. *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California, Berkeley.
10. Streeck, J. 1996. A little Ilokano grammar as it appears in interaction. *Journal of Pragmatics*, 26:189–213.
11. Verhaar, J. 1997. *Toward a Reference Grammar of TOK PISIN. An Experiment in Corpus Linguistics*. Oceanic Linguistics Special Publication No. 26, University of Hawai‘i Press, Honolulu.