Proceedings of Disfluency in Spontaneous Speech Workshop, Berkeley, California, 1 July 1999, pp. 3–6.

A COMPARATIVE ANALYSIS OF DISFLUENCIES IN FOUR SWEDISH TRAVEL DIALOGUE CORPORA

Robert Eklund
Telia Research AB, Farsta, Sweden

ABSTRACT

This paper reports on ongoing work on disfluencies carried out at Telia Research AB. Four travel dialogue corpora are described: human–“machine”–human (Wizard-of-Oz); human–“machine” (Wizard-of-Oz); human–human and human–machine. The data collection methods are outlined and their possible influence on the collected material is discussed. An annotation scheme for disfluency labelling is described. Preliminary results on five different kinds of disfluencies are presented: filled and unfilled pauses, prolonged segments, truncations and explicit editing terms.

1. INTRODUCTION

Current automatic speech recognition (ASR) and human–computer dialogue systems have attained a technological level that allows use in every-day commercial applications, as long as the tasks are suffiently constrained. In order to allow more open-ended speech input, certain phenomena typical of spontaneous speech need to be modelled. One such phenomenon is the processing of disfluencies (pauses, truncations, prolongations, repetitions, false starts etc.), or DFs for short, where more basic knowledge is needed in order to acquire more accurate modelling of spontaneous speech. To obtain such basic knowledge, a first necessary step is thus to study DFs in application-like situations. This paper describes ongoing work at Telia Research AB, where DFs in the travel booking domain are studied.

2. METHOD

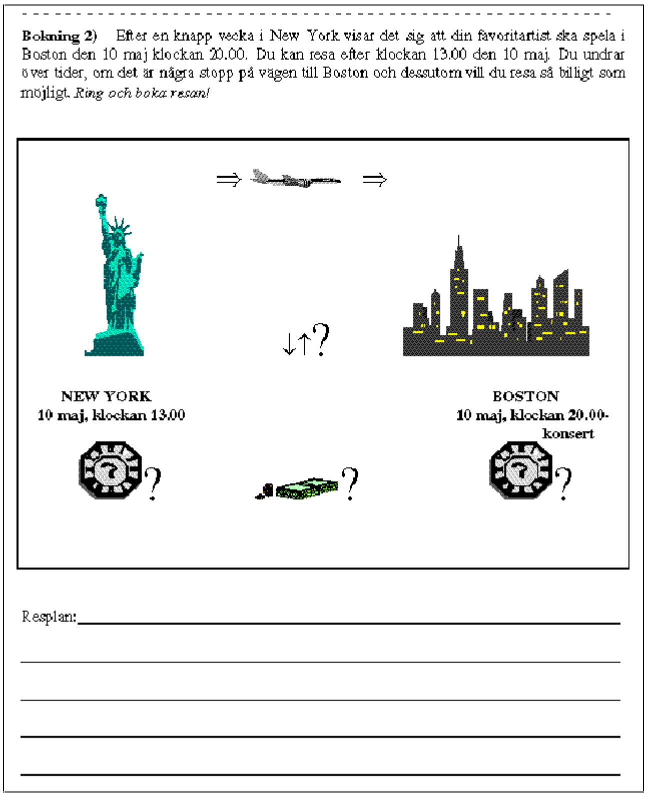
Travel booking dialogues in four different corpora are studied. The data were collected as a part of the Spoken Language Translator (SLT) project at Telia Research AB [1]. Since the SLT project work was carried out within the Air Travel Information Service (ATIS) domain [4], Swedish ATIS data were collected for language modelling and recognizer training purposes.

2.1. General Set-up and Subjects

The bookings were made over a telephone line, but high quality recordings were also made in order to facilitate acoustic analyses. The subjects were all Telia employees and were used to business travel bookings. As far as possible, the subjects were balanced over gender and age.

2.2. Corpus 1: WOZ–1 / Human–“Machine”–Human

In order to avoid too strong a colouring effect from instructions, the tasks were given in a mixture of written and pictorial form (cf. Figure 1). A more detailed description of this data collection session is found in [6].



Bokning 2) Efter en knapp vecka i New York visar det sig att din favoritartist ska spela i Boston den 10 maj klockan 20.00. Du kan resa efter klockan 13.00 den 10 maj. Du undrar över tider, om det är några stopp på vägen till Boston och dessutom vill du resa så billigt som möjligt. *Ring och boka resan!*

⇒ ⇒
↓↑?
NEW YORK
10 maj, klockan 13.00
BOSTON
10 maj, klockan 20.00-
konsert
?
?
?

Resplan:

Figure 1: Task sheet number two (of ten) for the WOZ–1 corpus. The text reads: “After almost a week in New York you are told that your favourite artist is going to perform in Boston on May 10, at eight o’clock in the evening. You can leave after 1 p.m. on May 10. You want to know the times, whether there are any stopovers on the way to Boston, and moreover, you want to travel as inexpensively as possible. *Make the call and do the booking!*” The subjects were allowed to scribble down supporting notes prior to the call.

2.3. Corpus 2: WOZ–2 / Human–“Machine”

During a later phase in the SLT database (SLT-DB) project, it was decided to expand the domain from ATIS to business travel bookings. This meant that data encompassing hotel information, car rental and so on were needed. Thus, the goal was here to record business travel dialogues between users and a simulated database application. The user was given the tasks in almost exlusively pictorial form (cf. Figure 2). A more detailed description is found in [5].

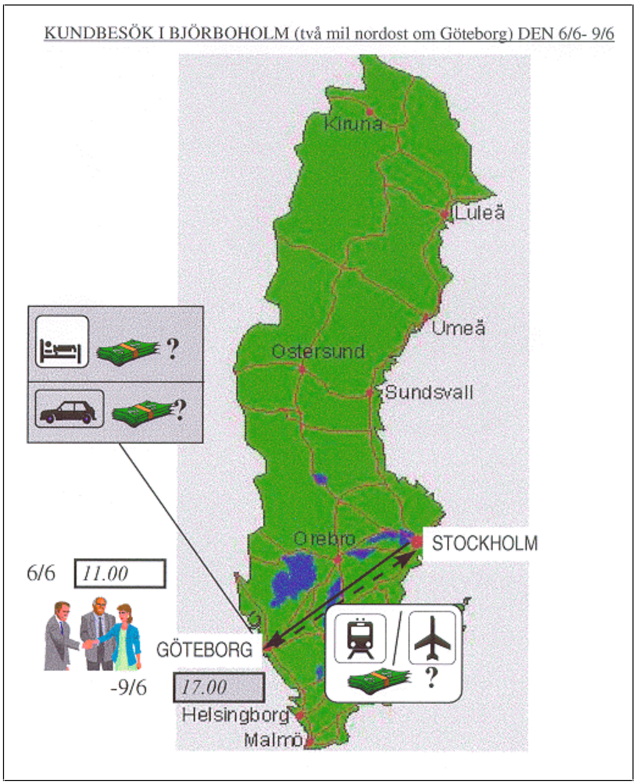


Figure 2: Task sheet number one (of three) for the WOZ–2, Nymans and Bionic corpora. The header reads: "Client visit in Björboholm (twenty kilometers north-east of Gothenburg) between June 6 and June 9." Arrows indicate departure and arrival, and icons and question marks indicate that inquiries should be made concerning the prices of trains and flights, accomodation and car rental. On this task sheet, dates and hours are given in figure format, whereas on other task sheets, icons of calendars and clocks were used.

2.4. Corpus 3: Nymans / Human–Human

In order to get a better grasp of general linguistic phenomena (grammatical, disfluencies, prosodic etc.), it was decided to collect authentic human–human dialogues. The subjects were given the same tasks as in WOZ-1 (cf. Figure 2).

2.5. Corpus 4: Bionic / Human–Machine

The goal of this data collection was to obtain authentic human–machine data. The word 'bionic' is used since authentic components were used to the extent that it was technically feasible. Since the recognizer did not cover some Swedish cities, a wizard was used to simulate recognition. Once again, the subjects were given the same instructions as were used in WOZ-2 and Nymans (cf. Figure 2).

2.6. Summary

The data thus collected are summarized in Table 1. All corpora were first transcribed orthographically by a transcription agency, but are presently being re-transcribed by the author to adhere to the annotation system described in section 3.

Table 1: Summary Statistics. The top row figures are based on the transcriptions made by the the transcription agency and are given to indicate the full size of the corpora. The bottom row (and slightly smaller) figures are based on the retranscribed and labelled material. Notes: ♠ = Only 23 subjects are labelled so far. * = Not fully transcribed. † = Note that there are eight subjects and two travel agents. ‡ = Only the subjects are transcribed and labelled, not the travel agents.

	WOZ–1	WOZ–2	Nymans	Bionic
Method	Script	Picture	Picture	Picture
No. subjects	39	47♠	10†	16
Male/Female	19/20	31/16	7/3	9/7
No. dialogues	390	131	24	*
Labelled	79	55	20‡	16
No. utterances	3,722	3,602	2,899	*
Labelled	957	1,632	1,323‡	517
No. words	29,645	27,277	21,611	*
Labelled	6,181	12,142	7,159‡	3,117

In all corpora there are a large number of one-word utterances, e.g. confirming utterances like "ja" (yes) and so on. Since one-word utterances are less likely to be disfluent, the relative rates of these affect the number-of-DFs/number-of-utterances ratio. There is reason to believe that one-word utterances are more common in Nymans than in the other corpora, since conversation support is more likely to occur in human–human conversation. However, we have presently not developed a reliable way of excluding one-word utterances, so all figures given include one-word utterances.

3. ANNOTATION

All corpora are labelled according to an annotation scheme first presented in [2]. This system is based on the annotation scheme developed by Shriberg [7], with some extensions and minor changes. The biggest differences are the explicit labelling of prolonged segments and that durations are explicitly given for filled and unfilled pauses, as well as for prolongations. The rationale of this system is two-fold. First, the aim was to use a system that could be easily mapped between languages, in order to facilitate cross-linguitic comparisons. Preliminary results from a cross-language study of Swedish and American English are presented in [3]. Second, since the annotation scheme in Shriberg is pre-theoretical, it could be assumed that it is easily portable other languages. Although Swedish is typologically very close to English, using a method similar to that of Shriberg could serve as test case for the generality of the labelling method. If the labelling method is indeed portable between languages, we would be one step closer toward a general tool and method for disfluency labelling across languages, which is a desideratum in the quest for deeper knowledge of human speech production.

3.1. Basic Annotation Scheme

The following disfluency categories are described in this paper:

`f<` ... `>f` **Filled pause (FP)** Marks the beginning and end of filled pauses, most often realised as "eh" or "öh" in Swedish. The filler word is written between the `f<` ... `>f` markers.

`u<` ... `>u` **Unfilled pause (UP)** Marks the beginning and end of unfilled pauses, i.e., silence. If heavy inhalations, exhalations or other similar phenomena occur during an unfilled pause, they are labelled between the `u<` ... `>u` markers.

`p<` ... `>p` **Prolongation (PR)** Marks prolonged segments. Hesitations can be realized as prolonged segments. i.e., a word might be pronounced with one or more segments markedly longer than in normal, fluent speech (e.g., "I want a fffffffflight to…"). The prolonged segment is indicated within curly brackets, and it is also indicated whether it is word-initial, word-medial or word-final. For instance the label **`{f-}`** indicates that the prolonged segment is a word-initial [f]. Swedish makes extensive use of productive compounding, and word boundaries within compounds are indicated using a hash sign # .

`e` **Explicit Editing Term (EET)** Words like "Sorry", "No, wrong", "I mean" and so on. EETs can be labelled in two ways: Either each word is counted, e.g. "Sorry" = 1; "I'm sorry" = 2 (or 3), or each "EET unit" is counted "I'm sorry" = 1. Although there are arguments in favour of the latter method — DFs rarely occur within such "packages" — we have opted for the former in this paper.

`/` **Truncation (TR)** Marks an interrupted word, either in repairs or caused by an intervening system or agent. The former, i.e., self-induced truncations, are of greater interest for our purposes, but the distinction is not made in this paper.

3.2. Repairs

Repairs are also labelled but both the labelling and the analyses of repairs are in a preliminary phase and a full discussion of these topics will have to await further work.

4. RESULTS AND ANALYSES

In this paper, we focus on FPs, UPs, PRs, EETs and TRs. The reason for choosing these types of DFs is not that they necessarily form a natural group, but rather to provide a general description of the material. Moreover, there might be more or less strong interaction between some of these types (e.g., FPs are often followed by UPs), and there is surely interaction between these types and repairs, which we have not looked at yet. The overall rates of the said DFs are shown in Table 2.

4.1. General observations

We first summarize some general observations on FPs, UPs, EETs and TRs. Since PRs is the hitherto least described DF type, these are treated in section 4.3.

4.1.1. Overall differences The first observation to be made is that the Bionic corpus exhibits the largest number of disfluencies, except for EETs. In fact, the ratio of the number of the included DFs and the number of labelled utterances is higher than 1, which indicates that there is a higher-than-100% chance that an arbitrary utterance will be disfluent in this corpus. Since the tasks were the same for WOZ–2, Nymans and the Bionic corpora, we may assume that this difference is not due to task details. It may be the case that the actual quality of the feedback—simulated synthesis in WOZ–1 and WOZ–2, human being in Nymans and real synthesis in Bionic—is the decisive factor. Even if the actors portraying the synthesizer spoke in a monotone voice and used a fixed set of standardized utterances, they did not exhibit jitter and similar phenomena associated with a real synthesizer. From this follows that one could possibly expect a larger number of disfluencies in real applications than in WOZ simulations.

Table 2: Summary of DF Rates. For each corpus, the percentages for the labelled material (as indicated in Table 1) are given, broken down by DF type. The figure is then divided by the number of labelled utterances and words in the corpus (multiplied by 100 to give percentages).

	WOZ–1	WOZ–2	Nymans	Bionic
Total no. FPs	225	400	182	145
No. FPs / no. utts	23.5	24.5	13.7	28.0
No. FPs / no. words	3.6	3.2	2.5	4.6
Total no. UPs	440	990	435	370
No. UPs / no. utts	46.0	60.7	32.9	71.6
No. UPs / no. words	7.1	8.1	6.0	11.9
Total no. PRs	94	84	107	67
No. PRs / no. utts	9.8	5.1	8.1	13.0
No. PRs / no. words	1.5	0.7	1.5	2.1
Total no. EET	20	42	10	10
No. EETs / no. utts	2.0	2.5	0.7	1.9
No. EETs / no. words	0.3	0.3	0.1	0.3
Total no. TRs	57	74	213	70
No. TRs / no. utts	6.0	4.5	16.1	13.5
No. TRs / no. words	0.9	0.6	3.0	2.2
Σ of included DFs	836	1,590	947	662
No. DFs / no. utts	87.4	97.4	71.6	128.0
No. DFs / no. words	13.5	13.1	13.2	21.2

4.1.2. Filled and Unfilled Pauses FPs and UPs are far more common than the other types of DFs, and UPs are generally twice as common as FPs.

4.1.3. Explicit Editing Terms and Truncations A somewhat surprising result is the low number of EETs in Nymans. One would assume that EETs were more common in human–human communication than to an inanimate system, but this is not the case. (Once again, the Bionic corpus is slightly different.) One possible explanation, however, could be related to the larger number of TRs; in Nymans the subjects are more often interrupted, and thus do not have either the opportunity (or reason) to put in EETs, since the experienced travel agent reacts to inconsistent information with a question. It must be pointed out that EETs are rare in all the corpora.

4.2. Durational DFs

In the set of DFs we look at in this paper, FPs, UPs and PRs stand out from EETs and TRs by being durational in nature, i.e., an FP, UP or TR can be stretched in time. Preliminary durational observations for these three types of DFs are presented in Table 3. The general tendency is that for all DFs and corpora, the pattern is PRs < FPs < UPs. Two things need be pointed out. First, whereas UPs can be stretched to great lengths, this does not occur for FPs or PRs. This explains the high standard deviation for UPs in the WOZ–2 corpus, where extreme outlier values occur, which was possible since the wizards were not instructed to take the initiative after a specified number of seconds.

Table 3: Mean Durations (in seconds) for FPs, UPs and PRs. Standard deviations are given in small figures.

	Mean FP	Mean UP	Mean PR
WOZ–1	0.53	0,62	0,25
sd	0,28	0,52	0,16
WOZ–2	0,44	1,01	0,30
sd	0,21	2,57	0,15
Nymans	0,49	0,54	0,29
sd	0,24	0,62	0,12
Bionic	0,43	0,62	0,25
sd	0,24	0,62	0,15
All	0,48	0,60	0,27
sd	0,25	0,74	0,14

Second, a distinction must be made between short UPs and very clear, computer-directed speech, where each word is uttered in isolation, but in a fluent manner. We are here dealing with two distinct speech styles. The durational values for FPs, UPs and PRs are fairly stable across the corpora, which leads us to believe that they are not that sensitive to either task details or general settings, but rather originate at lower levels in the speech production system.

4.3. Prolongations

There are basically two questions one can ask with regard to PRs: First, what kind of segments are prolonged? Second, what position in the word is favoured? Preliminary observations answering these questions are shown in Table 4.

Table 4: Phone Type and Position of Prolongations. For each corpus the percentages of phone position is given. Within each class, the percentages of phone class is given.

	WOZ–1	WOZ–2	Nymans	Bionic
% Initial phone	26.6	28.6	25.2	37.3
% vowel	1.0	0.0	14.9	8.0
% cons +sonorant	1.0	29.1	48.1	40.0
% cons –sonorant	24.6	70.9	37.0	52.0
% Medial phone	19.1	25.0	15.0	23.9
% vowel	9.6	38.1	25.0	31.2
% cons +sonorant	5.3	4.8	18.8	25.0
% cons –sonorant	4.2	57.1	56.2	43.8
% Final phone	54.2	46.4	59.8	38.8
% vowel	9.6	17.9	34.4	26.9
% cons +sonorant	30.8	71.8	53.1	61.8
% cons –sonorant	13.8	10.3	12.5	11.5

In the WOZ corpora and Nymans, initial, medial and final phone prolongations occur in the (roughly) 30–20–50 proportions mentioned in [3]. Once again, the Bionic corpus behaves slightly differently, however. Oddly enough, in no case is a vowel the preferred segment. This could depend on labelling, since there is a certain risk of judging prolonged vowels as “normal”. The most important observation, however, is that all kinds of segments can be prolonged, including voiceless stops. In fact, examples like “flyge……t” (the fligh……t) — where the occlusion phase is prolonged — are quite common. A general problem with labelling PRs occurs at the bottom end of the durational scale. While there are clear cases of very marked prolongation, it is quite often hard to say when a segment is prolonged slightly. However, since PRs undeniably exist, one would need to develop a method of labelling them consistently.

5. DISCUSSION

Although most of the material remains to be labelled and analyzed, some tendencies are clear. First, the data collection method and set-up clearly influences the material. Thus, it seems to be the case that the use of a real synthesizer in the Bionic corpus yielded a higher rate of DFs than the two WOZ collections. An interesting preliminary result, still needing corroboration, is that WOZ simulations appear to give results that are closer to human–human interactions than to human–machine interactions, underscoring the fact that human–human control data, in this case the Nymans corpus, are important for an accurate understanding of the processes involved in human–machine interaction. Second, PRs occur in all corpora in similar proportions, and arguably serve the same function as do FPs and UPs. Third, UPs and FPs are by far the most common DFs in all corpora, while EETs, TRs amd PRs are less frequent. A final point to be made is that it seems that Shriberg’s approach to DF labelling is indeed portable to Swedish.

ACKNOWLEDGMENTS

The set-up and desiderata of the data collection sessions were discussed and agreed upon between Catriona MacDermid, Camilla Eklund, Jaan Kaja, Mats Wirén and the author. Most of the work was carried out by Catriona and Camilla, with the assistance of Nils Meinhard. Thanks also to Carina Ekedahl and Lennart Svanfeldt at the travel agency Nyman & Schultz AB for participating in the human–human data collection. Also thanks to Eva Lindström and Anders Lindström for comments on draft versions of this article.

REFERENCES

[1] Becket, R., P. Boullion, H. Bratt, I. Bretan, D. Carter, V. Digalakis, R. Eklund, H. Franco, J. Kaja, M. Keegan, I. Lewin, B. Lyberg, D. Milward, L. Neumeyer, P. Price, M. Rayner, P. Sautermeister, F. Weng & M. Wirén. *Spoken Language Translator: Phase Two Report*. Telia Research AB and SRI International, 1997.

[2] Eklund, R. Interaction between prosody and discourse structure in a simulated man–machine dialogue. *Journal of the Acoustical Society of America*, Vol. 102, No. 5, Pt. 2, December 1997, 3202 [Abstract], 1997.

[3] Eklund, R. & E. Shriberg. Crosslinguistic Disfluency Modelling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues. *Proceedings of ICSLP 98*, Sydney, November 30–December 5, Paper 805, Vol. 6, 2631-2634. CD-ROM available from Causal Productions Pty Ltd, PO Box 100, info@causal.on.net, 1998.

[4] Hemphill., C.T., J.J. Godfrey & G.R. Doddington. The ATIS Spoken Language Systems Pilot Corpus. *Proceedings of DARPA Speech and Natural Language Workshop*, 96–101, 1997. `http://www.ldc.upenn.edu/readme_files/atis/sspcrd/corpus.html`

[5] MacDermid, C. & C. Eklund. *Report on the First WOZ Simulation for the SLT-DB Project*. Technical report, Telia Research AB, 1997.

[6] MacDermid, C. & C. Eklund. *Simulering av en automatiserad översättningstjänst för resebokningar.* Technical report, Telia Research AB, 1996.

[7] Shriberg, E. *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California, Berkeley, CA, 1994.