Proceedings Fonetik 2000, May 24–26, 2000, University of Skövde, Sweden, pp. 49–52.

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.

Robert Eklund
Telia Research AB, S–123 86 Farsta, Sweden

Abstract

This paper describes the collection of authentic human–human air travel information data in Tok Pisin, the pidgin/creole language spoken in Papua New Guinea. Pros and cons of authentic data are discussed, as compared to data collected in more controlled settings like Wizard-of-Oz simulations. Some unexpected real-life phenomena that affect the data, and normally do not occur in corpora compiled from Wizard-of-Oz simulations, are described.

Introduction

Current automatic speech recognition systems have gained a level of stability that makes possible inclusion in real-life applications, for example fully automatic dialogue systems. There are, however, some problems that are not yet satisfactory handled, like, among others, the occurrence of so-called disfluencies, typical of human spontaneous speech (cf. Shriberg 1994). In order to study, and model, the occurrence of disfluencies, various kinds of data are needed, covering, among other things, a variety of languages, different domains/services (air travel, hotel room bookings etc.), different kinds of dialogues, (human–human, human–machine etc.), different channels (face-to-face, telephone, screen-based) and so on. This paper describes the collection of human–human air travel information (ATIS) (Hemphill et al. 1990) dialogues in Tok Pisin, the pidgin/creole language spoken in Papua New Guinea (Verhaar 1995). The objective of the collection was to enable crosslinguistic studies of disfluencies, since very little has been done in this field (cf. Eklund & Shriberg 1998). The paper focuses on methodological aspects of the collection of authentic data.

Method

The dialogues were collected at the Kavieng Airport, New Ireland Province, Papua New Guinea in December 1999 and January 2000. When collecting authentic data, the first thing to do is to ask the permission of the participants. Consequently, prior to the recording sessions, the author presented himself at the Air Niugini sales office at the airport and explained to the staff the goal of the study. The sales agents kindly agreed to participate. When a customer arrived to either purchase or confirm a ticket, the author introduced himself as a linguist interested in Tok Pisin, and asked the permission to record the interaction between the customer and the travel agent. All subjects accepted to be recorded. A portable professional tape deck with a stereo microphone was used . The author either sat or stood — depending on what travel agent was involved — next to the customer in order to adjust recording volume and microphone direction (Plates 1a and 1b). After having completed their business with the travel agent, the author asked the subjects where they were born, where they resided and what their native language was, besides Tok Pisin. The subjects — or, when applicable, their children — were also given small, symbolic gifts, e.g. “thank you” postcards with pictures of typical Swedish phenomena (e.g. snow, elks and the like).

PLATE 2a. Air Niugini sales office, Kavieng Airport, New Ireland, Papua New Guinea. Travel agent Loris Levi and subject 03. (Photograph by Robert Eklund.)

PLATE 2b. Air Niugini sales office, Kavieng Airport, New Ireland, Papua New Guinea. Travel agent Nianne Kelep, subject 39 and the author. (Photograph by Eva Lindström.)

The Corpus

In this way, 39 dialogues were collected, and are presently being digitized on disk and are transcribed and analyzed according to the scheme presented in Eklund (1999). The dialogues vary in length from very short flight confirmations to long, full-fledged purchases of tickets.

Discussion

The dialogues mentioned in this paper cover a language hitherto unexplored in several of the dimensions listed above, e.g. task, domain and channel. The pros and cons of the way in which this corpus of authentic dialogues has been compiled are discussed in the following.

Data collection: General remarks

There are several ways to obtain data, ranging from *(a)* introspection; *(b)* laboratory speech; *(c)* 'Wizard-of-Oz' (WOZ) simulations; *(d)* collecting authentic data, either without prior permission (e.g. by tapping a telephone line) or with prior permission (as in this study); *(e)* translation of material obtained by using *(a)* – *(d)*. There are advantages and disadvantages associated with all of these methods (cf. Bretan, Eklund & MacDermid 1996). For instance, *(a)* is a very inexpensive method, but is likely to provide very limited data. While method *(b)* is very nice with regard to sound quality, the material collected is more or less inevitably dependent on *(a)*, and hence results in not-so-natural linguistic data. WOZ simulations have been proven to provide more spontaneous data, but are still not fully natural, since the tasks given to the subjects are not real-life tasks that the subjects need to carry out, but might be seen as "tests" by the subjects. Furthermore, WOZ data quite often contain meta comments, e.g. to the session leader, indicating an awareness of the presence of another person, although that person is not in the same room. Moreover, the subjects might misunderstand the objective of the study, and either consciously or subconsciously adapt their speech to whatever need they might think the session leader might have, or simply just believe that they are being tested, which leads to a degree of nervousness that would not be there in a fully authentic situation (cf. Eklund & Shriberg 1998).

The present corpus: advantages

The main, obvious, advantage of this corpus is that all tasks and dialogues are authentic. The general impression was also that the subjects were not intimidated by the presence of the author (and his gear), although full certainty in this domain can never be obtained. Although some of the interactions were short (simply confirming previous bookings), a fair amount of longer, mixed-initiative, interactions were collected.

The present corpus: disadvantages

A few problems came up that the author had not thought of beforehand, some of which are likely to change the nature of the data, whereas others will make the data difficult to analyze, or simply useless. *(1)* The sales office was located at the airport, which meant that some dialogues were disturbed by jet aircraft, taxing in just a few meters from the microphone. *(2)* A more general problem, not encountered in "staged" situations such as WOZ simulations, was the fact that several subjects did not turn up alone, but with their children, spouses, friends or colleagues, which meant that the "dialogues" did not all turn out to be two-part, since the subjects intermittently spoke to the accompanying person(s), who quite often also took part in the interaction. *(3)* During the pauses when the travel agent spent time looking for flights in her computer system, the subjects would often initiate a dialogue with the author, very much indicating an awareness of his presence. Whether or not this affected the interaction with the travel agent is impossible to say, and the alternative to stay invisible and have the travel agents record the dialogues would have meant that the subjects would not have seen/known *who* would use the recordings, which probably could have made them feel uneasy about the presence of a microphone. Thus, a session leader with whom they could casually chat was deemed a better way to carry out the data collection. *(4)* The timing was not optimal, since late December through January is a period where a lot of customers just confirm flights, rather than make full bookings.

Conclusions and future work

Whatever method is chosen to collect linguistic data is bound to affect the data *per se*, and there is often a trade-off between what you gain in one dimension, e.g. naturalness, and what you lose in another, e.g. sound quality. Of the observations mentioned in this paper, the most interesting is perhaps that in the development of dialogue systems, it is rarely — if ever — considered that the customer might not be alone, and that comments from both the travel agent and the customer might be "outside" the task proper. Also, as shown in point *(4)*, picking the time can seriously affect the data collected. As far as future work goes, the corpus described in this paper will be analyzed with regard to disfluency occurrence, which hopefully will provide further insights with regard to universal tendencies in this domain. It will also be used to study the differences between different ways of collecting data, and the resulting linguistic consequences, since it can be compared with the compilation of the Tok Pisin ATIS corpus described in Eklund (1998), where translation (of WOZ data) and elicitation were used.

Acknowledgements

I would like to thank the Air Niugini travel agents at Kavieng Airport, Loris Levi, Nianne Kelep and Liza Gabriel, for helpful assistance. Thanks to Jenny Xomerang for helpful tips. Thanks to Eva Lindström for initial contacts, Tok Pisin tuition and for inviting me to New Ireland. I would also like to thank Dicks Thomas for Tok Pisin advice. Thanks to Pär Lindström for providing the tape recorder. Thanks to Anders Lindström for comments.

References

Bretan, Ivan, Robert Eklund & Catriona MacDermid. 1996. Approaches to Gathering Realistic Training Data for Speech Translation Systems. *Proc. IVTTA – 1996 IEEE Third Workshop, Interactive Voice Technology Telecommunications Applications*, September 30–October 1, 1996, Basking Ridge, New Jersey, pp. 97–100.

Eklund, Robert. 1999. A Comparative Study of Disfluencies in Four Swedish Travel Dialogue Corpora. *Proc. of Disfluency in Spontaneous Speech Workshop*, Berkeley, California, 1 July 1999, pp. 3–6.

Eklund, Robert. 1998. "Ko Tok Ples Ensin bilong Tok Pisin" or the TP-CLE: A first report from a pilot speech-to-speech translation project from Swedish to Tok Pisin. *Proc. ICSLP 98*, Sydney, November 30–December 5. Paper 804, Vol. 4, pp. 1131–1134. CD-ROM från Causal Productions Pty Ltd, PO Box 100, mailto:info@causal.on.net.

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

Hemphill, Charles T., John J. Godfrey & George R. Doddington. 1990. The ATIS Spoken Language Systems pilot corpus. *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, PA., pp. 96–101.
`http://www.ldc.upenn.edu/readme_files/atis/sspcrd/corpus.html`

Shriberg, Elizabeth. 1994. *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California, Berkeley, CA.

Verhaar, John. 1995. *Toward a Reference Grammar of TOK PISIN. An Experiment in Corpus Linguistics*. Oceanic Linguistics Special Publication No. 26., University of Hawai'i Press, Honolulu, 1995.