"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

Robert Eklund

Telia Research AB, Farsta, Sweden

ABSTRACT

This paper describes an operational speech-to-speech translation system from Swedish to Tok Pisin within the framework of the Spoken Language Translator project, SLT [1]. The domain of translation is ATIS [11]. The grammar formalism used in the SLT project is the Core Language Engine, CLE [2]. A general presentation of Tok Pisin is provided, as well as a description of some grammatical characteristics of Tok Pisin of potential interest for the testing of grammar machines. The first step of a CLE implementation of Tok Pisin is described. A corpus of Tok Pisin ATIS data has been created from data collected on location in New Ireland, Papua New Guinea, and observations are made as to the relative importance of some of the grammatical phenomena discussed in the paper. A Tok Pisin synthesizer based on an already existing Swedish concatenative synthesis is described. Despite a marked Swedish accent, preliminary evaluation indicates that intelligible speech output is produced.

1. INTRODUCTION

Machine translation, or MT for short, nowadays covers many languages, such as English, French, Japanese, German, Korean and Swedish. To the best of our knowledge, no MT project has so far included a pidgin or creole language. This paper constitutes the first report from a project of automatic speech-to-speech translation from Swedish to Tok Pisin.

One can hardly claim that there is an immediate commercial interest in automatic Swedish-to-Tok Pisin speech translation. However, Tok Pisin exhibits some traits that are of practical and theoretical interest for the testing of grammar formalisms.

Although pidgin languages are often considered 'simple' (e.g. due to their lack of inflectional morphology), they exhibit traits not commonly found in languages normally included in machine translation projects. This makes them potentially interesting for the testing of MT machines.

Grammatical descriptions of Tok Pisin exist [10, 12, 14, 15, 16, 17], as well as works on translation to and from Tok Pisin [5, 7], but there have so far (to the best of our knowledge) been no descriptions of Tok Pisin from an MT perspective. This paper points to some of the features of Tok Pisin that could be of interest for the testing of grammar formalisms.

However, since it is not obvious that all linguistic phenomena to be described are of equal importance within a restricted domain, the relative importance of the said phenomena is investigated. This discussion is based on observations made on a corpus of authentic linguistic data.

2. THE SLT PROJECT

The general framework of this pilot project is the Spoken Language Translator [1], or SLT for short, a speech-to-speech translation project whose main focus so far has been English [1], Swedish [1, 8, 9] and French [13]. It is a joint project between Telia Research AB and SRI International. The domain of the project is the Air Travel Information Service, ATIS [11]. As is the case in all speech translation systems, translation occurs in three steps: speech recognition, text translation, and speech synthesis. The three modules of the SLT project are briefly described in the following passages.

2.1. The Recognizer: SRI Decipher

The recognizer of the project is the SRI speech recognizer Decipher®. It was trained for Swedish during the second phase of the SLT project [3].

2.2. The Core Language Engine

The grammatical engine is the Core Language Engine [2], a formalism developed by SRI Cambridge. It is a unification-based formalism aiming to be theory-neutral. The CLE maps between natural sentences and logical representations of their meaning. Every sentence is given both a syntactic and a corresponding semantic interpretation. The first modules in the analysis chain (lexical, morphological, syntactic and semantic) output a set of *quasi logical forms* (QLFs), where certain aspects are left without analysis, such as quantifier scope and anaphor resolution, which are handled at later stages in the process. Translation is transfer-based, and occurs at QLF level, but there is also a set of unidirectional word-to-word rules, used as a fallback method, in case QLF analysis fails.

2.3. Speech Synthesis

The Swedish synthesizer has been developed at Telia Research AB. It is a concatenation synthesis that uses mainly demi-syllables. Currently some 15,000 units have been recorded. A more detailed description of the synthesizer is given in [6].

3. 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 with nearly 800 languages. (The other two official languages are English and Hiri Motu.) Although around 80 % of the lexicon is derived from English (figures vary), the syntax is predominantly Austronesian. The basic word order is SVO.

3.1 Some Grammatical Traits of Tok Pisin

The Austronesian substratum in Tok Pisin appears on several levels in the syntax. A few traits of interest will be described in the following passages.

Distinction between inclusive and exclusive personal plural pronouns. Tok Pisin discriminates between e.g. **mipela** ("we", exluding the addressee) and **yumi** ("we", including the addressee).

Predicate marker. Tok Pisin makes use of a predicate marker **i**, which precedes the predicate in cases where the subject contains some kind of third-person element. (This means that it does not appear after some subject pronouns.) Although it is sometimes seen written together with the predicate verb (e.g. in the weekly newspaper WANTOK), and sometimes with the preceding pronoun (ditto), it definitely exists in its own right, and marks the entire predicate. For example, in negated predicates, the negation **no** goes between the predicate marker and the predicate verb.

Serial verb constructions. All verbs of movement or direction are serial verb constructions in Tok Pisin, specifying whether the direction implied is towards (**i kam**) or away from (**i go**) the speaker.

Aspect/tense marking. Tok Pisin encodes aspect (continuous and completed) and tense (mainly future) by means of free-standing markers with a relatively free distribution. The continuous marker is **i stap** or **wok long** and the completive marker is **pinis**. The future marker is **bai**, which is placed either in sentence-initial or predicate-initial position. This means that the subject pronoun is (often) placed in between the future marker and the predicate verb.

Reduplicative morphology. Although Tok Pisin, like pidgin and creole languages in general, lacks inflectional morphology, reduplicative morphology is very productive, and serves many purposes. Since reduplication affects the lexicon in a very profound way, it could potentially provide a challenge to the CLE morphology module.

3.2. Relative Importance of Traits

Of the traits listed above, it was surmised that some of the phenomena are more crucial than others within a specific, restricted domain, such as ATIS. For example, reduplicative morphology is mainly an expressive tool and is not likely to show up in a fairly formalized booking situation. On the other hand, verbs of movement are very much a part of a travel booking situation. To investigate the relative importance of the linguistic phenomena above, domain-specific data were needed.

4. A TOK PISIN ATIS CORPUS

ATIS data were collected on location in Bimun village, New Ireland, Papua New Guinea. During the SLT project, different methods have been used to obtain ATIS data, and it has been shown that different methods to a large extent yield data of different quality [4]. Therefore, three different methods were employed .

4.1. Subjects

Data were collected from three subjects, all of whom were native speakers of Tok Pisin, and two of whom were experienced plane travellers.

4.2. Translations

The first method was simply to translate English ATIS sentences into Tok Pisin. The author orally gave the subjects typical ATIS sentences such as "I would like to go from Kavieng to Rabaul on Friday." The subjects then provided Tok Pisin translations, also in oral form, which were written down by the author.

4.3. Elicitation

In addition to translations, a second method was used in order to obtain data less biased by English. The author presented the subjects with situations, by saying things (in Tok Pisin) like: "Suppose you would like to know what the flights from Port Moresby to Madang are, what would you ask the travel agent?" The responses were then written down.

4.4. Simulated Booking

To obtain even more spontaneous data, a booking situation was simulated. The author provided some basic areas to be covered in a booking situation, in written and oral form, to a linguist fluent in Tok Pisin. She then impersonated a travel agent, and a native speaker of Tok Pisin "ordered" a flight ticket from her. Incidentally, the linguist had earlier participated in ATIS translations for the SLT project and was thus familiar with the purpose of the task. The author monitored the dialogue and wrote down the subject's utterances.

4.5. The Resulting Corpus

In this way, a small corpus of 169 Tok Pisin sentences was compiled, corresponding to 100 English "input" sentences. Despite the small amount of data thus collected, it was clear that the three different methods resulted in different data. An excerpt from the corpus is shown in Figure 1.

4.6. Linguistic Observations

As was hypothesized in 3.2., certain grammatical traits do not appear in the corpus. First, no instances of inclusive pronouns occur, which is not surprising, given that the travel agent most often is not included in the travel plans of the client. Second, only one instance of the completive aspect marker **pinis** occurs, which is also not surprising, since travel bookings are mainly a concern of the future, which makes completed actions rare. Third, all instances of reduplication could be considered lexicalized, and thus pose no problem to the morphology module.

However, grammatical phenomena that obviously call for attention are the future marker **bai**, which occurs several times in the data (in fifty percent of the cases with the subject in between **bai** and the predicate verb), and predicate marking with **i**, which is ubiquitous.

```
<e002>   How much does that flight cost?
<t002.c> Hamas dispela ron bilong balus i kost?
<t002.c> Wanem prais bilong dispela ron bilong
         balus?
<t002.c> Wanem pe bilong dispela ron bilong
         balus?
<t002.c> Hamas dispela flait i kost?
<t002.c> Wanem prais bilong dispela flait?
<t002.c> Wanem pe bilong dispela flait?
<t002.r> Hamas bilong baim tiket bilong dispela
         flait?
```

Figure 1: Excerpt from Tok Pisin ATIS corpus. The English "input" sentence (**.e**) and Tok Pisin translations (**.t**). (The function of the suffixes **.c** and **.r** is to distinguish between informants.) Some of the translations were the results of translation proper, some were elicited by described situations, and some were drawn from the enacted booking situation. In the case of the booking simulation, the English "original" was (re-)created after the Tok Pisin sentence was collected.

5. THE TP-CLE

At this stage, the CLE implementation of Tok Pisin is still rudimentary. On the grammar side, a bidirectional lexicon has been created, but this is currently not used in the translation process. On the translation side, so far only a small set of uni-directional word-to-word transfer rules have been implemented. An example of such a rule looks like this:

```
trule_ww([swe,tok],
   [flygning/nbar]
   >=
   [ron,bilong,balus]).
```

Note that phenomena like predicate marking could be given a provisory "solution", e.g. by writing rules that translate the Swedish preposition *till* ("to") into the composite Tok Pisin form **[i,go,long]**. This will produce a correct translation in the third person, but not in the first. However, since certain verbs tend to go with certain subjects, even simple rules like these can produce a surprisingly large amount of grammatically correct output. It goes without saying, however, that this is not a satisfactory solution to the challenges of Tok Pisin grammar.

6. THE SYNTHESIZER

In order to obtain full speech-to-speech translation, a Tok Pisin synthesizer was needed. Since the phoneme and demisyllable inventories of Tok Pisin are more or less properly included in the Swedish phoneme and demisyllable inventories, or can be approximated by phonetically similar items, a makeshift Tok Pisin synthesis was obtained by piggy-backing on the already existing Swedish synthesizer (cf. 2.3.).

The following five sentences, drawn from the Tok Pisin ATIS corpus, were synthesized and played to native speakers of Tok Pisin on location on New Ireland.

(1) **Mi laik painim ron bilong balus i kirap long Fraide.**
(I would like a flight that departs on Friday.) [0804_01.WAV]

(2) **Hamas bilong baim tiket i go long Mosbi?**
(How much does a ticket to Port Moresby cost?) [0804_02.WAV]

(3) **I gat sampela sit i stap long dispela ron bilong balus?**
(Are there any seats left on that flight?) [0804_03.WAV]

(4) **Raitim olgeta flait long Kavieng i go long Mosbi i gat stap long Manus na Madang.**
(List all flights from Kavieng to Port Moresby with stopovers in Manus and Madang.) [0804_04.WAV]

(5) **Wanem taim bai mi lusim dispela hap?**
(What time do I leave?) [0804_05.WAV]

Since the prosody rules are written for Swedish, the output has a strong Swedish accent, especially with regard to questions. However, the results of this rather informal testing showed that whereas sentence (1) was very hard to understand to most listeners, sentences (2) through (5) were understandable on both segmental and intonational levels, despite the strong accent.

7. TRANSLATION

The modules described in the previous passages are linked together and a small number of sentences spoken in Swedish are translated and spoken out in Tok Pisin. An example of a sentence that has received a grammatical translation is shown in Figure 2.

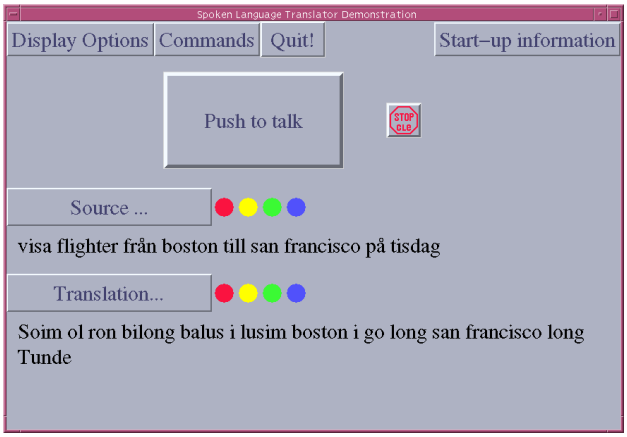

Figure 2: SLT demonstrator interface. Swedish input utterance and Tok Pisin translation. [0804_06.WAV] English translation: "Show flights from Boston to San Francisco on Tuesday."

8. EVALUATION AND FUTURE WORK

Although all the links in the translation chain currently have a makeshift or rudimentary character, a running system has been obtained. Due to the minimality of the system, an evaluation of the system was not attempted. Moreover, there was the problem of finding native speakers of Tok Pisin in Sweden to do the evaluation.

An ATIS corpus in Tok Pisin has been created, which, despite its small size, still provides information concerning a small, but

interesting, number of idiomatic expressions, and hints at the peripheral role of certain grammatical phenomena within this domain. However, it is clear that the corpus is too small to allow any far-reaching conclusions to be drawn. More data is clearly a desideratum.

The language module so far consists only of an implementation of a small set of word-to-word rules, covering only a few cases of grammatical translation. Clearly, the real linguistic interest lies in the implementation of a real Tok Pisin grammar and transfer rules at QLF level. This work is yet to be done.

The synthesizer has proven to be of some usability. However, its Swedish "heritage" clearly makes it sub-optimal. While segmental problems can only be solved by recording a native Tok Pisin speaker, much improvement could probably be obtained by writing Tok Pisin-specific prosody rules.

9. ACKNOWLEDGEMENTS

Whipping together a speech-to-speech translation system is obviously not a one-man feat. The author would like to thank the following people: For CLE competence and discussions, my SLT colleagues Ivan Bretan, Dave Carter, Manny Rayner and Mats Wirén; For help with the speech synthesis my colleagues Eva Öberg and Jaan Kaja; For help and assistance with regard to Tok Pisin, John Verhaar, Andrew Pawley, Christopher Stroud and Eva Lindström; For pidgin and creole language advice, Mikael Parkvall. Special thanks go to my informants in Bimun, New Ireland, Clemens (Kalamendy) Towil, Abraham Towil and Robert Sipa. Special thanks to Eva Lindström for playing the Tok Pisin speech synthesis to natives of New Ireland, for comments on draft versions of this article, and for impersonating a travel agent with such bravura.

10. REFERENCES

1. Agnäs, M.-S., Alshawi, H., Bretan, I., Carter, D., Ceder, K., Collins, M., Crouch, R., Digalakis, V., Ekholm, B., Gambäck, B., Kaja, J., Karlgren, J., Lyberg, B., Price, P., Pulman, S., Rayner, M., Samuelsson C. & Svensson, T. *Spoken Language Translator: First Year Report.* SRI Technical Report CRC-043, 1994.

2. Alshawi, H. (ed.) *The Core Language Engine*. MIT Press, 1992.

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

4. Bretan, I., Eklund, R. & MacDermid, C. Approaches to Gathering Realistic Training Data for Speech Translation Systems. *Proc. of IVTTA – 1996 IEEE Third Workshop, Interactive Voice Technology for Telecommunications Applications*, September 30 – October 1, Basking Ridge, New Jersey, 1996.

5. Conrad, B. Problems in translating from Tok Pisin to Mufian. In Verhaar (ed.*), Melanesian Pidgin and Tok Pisin. Proc. of the First International Conference of Pidgins and Creoles in Melanesia*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1990.

6. Eklund, R. & Lindström, A. How to Handle "Foreign" Sounds in Swedish Text-to-Speech Conversion. Approaching the 'Xenophone' Problem. *Proc. of the International Conference on Spoken Language Processing*, Sydney, November 30 – December 5, 1998. (Paper 514, these proceedings.)

7. Franklin, K.J. On the translation of official notices into Tok Pisin. In Verhaar (ed.), *Melanesian Pidgin and Tok Pisin. Proc. of the First International Conference of Pidgins and Creoles in Melanesia*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1990.

8. Gambäck, B. *Processing Swedish Sentences: A Unification-Based Grammar and Some Applications.* PhD Thesis. Swedish Institute of Computer Science, Kista, 1997.

9. Gambäck, B. & Rayner, M. The Swedish Core Language Engine. *Proc. 3rd Nordic Conference on Text Comprehension in Man and Machine*. Linköping, Sweden, 1992. (Also SRI Cambridge Technical Report CRC-025.)

10. Hall, R. *Melanesian Pidgin English. Grammar, Texts, Vocabulary*. Linguistic Society of America, Baltimore, MD, 1943.

11. Hemphill, C.T., Godfrey, J.J. & Doddington, G.R. The ATIS Spoken Language Systems pilot corpus. *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, PA., 1990.

12. Mihalic, J.J. *The Jacaranda Dictionary and Grammar of Melanesian Pidgin.* (Reprinted 1983.) The Jacaranda Press, Hong Kong, 1971.

13. Rayner, M., Carter, D. & Bouillon, P. Adapting the Core Language Engine to French and Spanish. *Proc. NLP-IA*, Moncton, New Brunswick, 1996. (Also SRI Technical Report CRC-060.)

14. Scorza, D. & Franklin, K.J. *An Advanced Course in Tok Pisin*. The Summer Institute of Linguistics, Ukarumpa, Papua New Guinea, 1989.

15. Verhaar, J. *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.

16. Wurm, S.A. & Mühlhäusler, P. *Handbook of Tok Pisin (New Guinea Pidgin*). Pacific Linguistics, Series C – No. 70. The Australian National University, Canberra, 1985.

17. Wurm, S.A. *New Guinea Highlands Pidgin: Course Materials*. Pacific Linguistics, Series D – No. 3, Australian National University, Canberra, 1971.