We1Q2

"Ko Tok Ples Ensin" or the TP-CLE: A first report from a pilot speech-to-speech translation project from Swedish to Tok Pisin

Robert Eklund
Robert.H.Eklund@telia.se

Telia Research AB
S-123 86 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.

Introduction and Rationale

Why Pidgins and Creoles in MT?

- Machine Translation, MT, nowadays covers many languages, such as English, French, Japanese, German, Korean, Swedish and others.
- To the best of our knowledge, no MT project has so far included a pidgin or creole language.
- However, pidgins/creoles exhibit some traits that are of practical and theoretical interest for the testing of grammar formalisms.
- Pidgin languages are generally considered 'simple' e.g. due to their lack of inflectional morphology.
- They do however exhibit grammatical trait not commonly found in languages normally included in machine translation projects.
- This paper constitutes the first report from a project of automatic speech-to-speech translation from Swedish to the pidgin/creole language Tok Pisin.
- Some of the linguistic phenomena of Tok Pisin that could be of interest for the testing of grammar formalisms are investigated, as well as their relative importance within a restricted domain.

Tok Pisin: A Brief Description

- Tok Pisin is an English-lexicon pidgin/creole language spoken in Papua New Guinea (see map page 3).
- It is one of the three official languages of Papua New Guinea, a nation with more than 800 languages.
- The other two official languages are English and Hiri Motu.
- Number of speakers, approx. 2 million.
- The syntax is predominantly Austronesian.
- The basic word order is SVO.
- Grammatical descriptions of Tok Pisin exist [10, 12, 14, 15, 16, 17].
- Works on translation to and from Tok Pisin also exist [5, 7].

Papua New Guinea

The Spoken Language Translator Project

Speech-to-Speech Translation

- The general framwork is the Spoken Language Translator, SLT [1].
- Joint project between Telia Research AB and SRI International.
- Speech-to-speech translation.
- Main focus so far Swedish [1, 8, 9] , English [1], and French [13].
- Also rudimentary versions for Spanish, Dutch, German, Korean, Danish.
- The domain is Air Travel Information Service, ATIS [11].

The Recognizer: SRI Decipher

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

The Core Language Engine

- The grammatical engine is the Core Language Engine, CLE [2], developed by SRI Cambridge.
- The CLE is a unification-based formalism.
- Aiming to be theory-neutral.
- Maps between natural sentences and logical representations of their meaning.
- Every sentence is given a syntactic and corresponding semantic interpretation.
- The first modules in the analysis chain output a set of quasi-logical forms, QLFs.
- Translation is transfer-based, but there is also a set of word-to-word rules used as a fallback method.

Speech Synthesis

- The Swedish synthesizer is developed at Telia Research AB [6].
- Concatenation synthesis using a female voice.
- Demisyllables with some additions.
- Around 15,000 units.

Tok Pisin: Some Grammatical Traits

Introduction

- The Austronesian substratum of Tok Pisin appears on several layers in the syntax.
- Some of these traits are briefly described below.

Inclusive/Exclusive Plural Pronouns

- Tok Pisin discriminates between *mipela* (we, excluding the addressee) and *yumi* (we, including the addressee.

Predicate Marking

- Tok Pisin makes use of a predicate marker *i*, which precedes the predicate in cases where the predicate contains some kind of third-person element. (This means that it does not appear after some personal pronouns.)
- In negated predicates, the negation *no* goes between the predicate marker and the predicate verb.

Serial Verb Constructions

- All verbs of motion or direction are serical verb constructions, specifying whether the direction implied is towards (*i kam*) or away (*i go*) from the speaker.

Aspect/Tense Marking

- Aspect (continuous and completed) are encoded by free-standing markers with a relatively free distribution.
- The continuous marker is *i stap* or *wok long* and the completed marker is *pinis*.
- The future marker is *bai*, which is placed either in sentence- or in predicate-initial position.

Reduplicative Morphology

- Reduplicative morphology is very productive.
- Could potentially provide a challenge to the lexicon.

A Tok Pisin ATIS Corpus (1)

Hypothesis

- Of the traits listed above, it was surmised that some are more crucial than others within a restricted domain such as ATIS, e.g., reduplication is mainly an expressive tool and is not likely to show up in booking contexts, whereas verbs of motions are ubiquitous.
- To investigate the relative importance of these traits, a corpus of ATIS data was compiled.
- The data were collected in Bimun village, New Ireland (top right corner of the map).
- Since different data collection methods yield different data [4], three different methods were used, described to the right.

Subjects

- Data were collected from three different subjects.
- All were native speakers of Tok Pisin.
- Two subjects were experienced plane travellers.

Translations

- The author orally provided the subjects with typical ATIS sentences (in English).
- The subjects provided Tok Pisin translations in oral form, which the author wrote down.
- Alternative translations were encouraged.

Elicitations

- The author presented the subject s 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.

Simulated Booking

- A booking situation was simulated.
- A linguist fluent in Tok Pisin impersonated a travel agent, and a native speaker ordered a flight ticket from her.
- The author monitored the dialogue and wrote down the subject s utterances.

A Tok Pisin ATIS Corpus (2)

The Resulting Corpus

- A corpus of 169 Tok Pisin sentences was compiled, corresponding to 100 English "input" sentences.
- Despite the small amount of data collected, it was clear that the three different methods resulted in different data.

Linguistic Observations

- As was hypothesized, certain of the listed grammatical traits do not appear in the corpus.
- No instances of inclusive plural pronouns occur, which is not surprising, since the travel agent most often is not included in the travel plans of the client.
- Only one instance of the completed aspect marker *pinis* occurs, which is also not surprising since travel bookings are a concern of the future, making completed actions rare.
- All encountered instances of reduplications could be considered lexical, and pose no problem to the lexicon.
- However, phenomena that call for attention are the future marker *bai*, and predicate marking, that are ubiquitous.

Tok Pisin ATIS Corpus: Excerpt

<e009> I want a one-way ticket to Madang.
<t009.c> Mi laikim wanpela we tiket i go long Mosbi.
<t009.r> Mi laikim wanwe tiket i go long Madang.

<e010> I would like a riturn ticket to Port Moresby.
<t010.c> Mi laikim riten tiket i go long Mosbi.
<t010.r> Mi laikim riten tiket i go long Mosbi.

<e011> I want a one-way ticket on Air Niugini.
<t012.c> Mi laikim wanwe tiket long Air Niugini.
<t013.c> Mi laikim wanwe tiket wantaim Air Niugini.
<t014.r> Mi laikim wanwe tiket long Air Niugini.

<e015> I want a return ticket on Air Niugini.
<t015.c> Mi laikim riten tiket long Air Niugini.
<t015.c> Mi laikim riten tiket wantaim Air Niugini.
<t015.r> Mi laikim riten tiket long Air Niugini.

<e016> I want the cheapest ticket.
<t016.c> Mi laikim tiket i sip.
<t016.r> Mi laikim tiket prais i daun.

The CLE Implementation of Tok Pisin

Grammar

- At the present stage, the CLE implementation of Tok Pisin is still very rudimentary.
- A bidirectional lexicon has been created, but this is currently not used in the translation.

Transfer Rules

- A small set of unidirectional word-to-word transfer rules have been created.
- An example of a rule looks thus:

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

- Phenomena like predicate marking could be given a provisory solution by writing rules that map single words into composite forms, such as the following rule for the preposition *till* (to):

```
trule_ww([swe,tok],
         [till/p]
         >=
         [i,go,long]).
```

- This would produce a correct a correct translation in the third person but not in the first person.
- However, since certain verbs tend to go with certain subjects, even simple rules like the one above could yield a fair amount of correct output.
- It goes without saying that this is not a satisfactory solution to the challenges of Tok Pisin grammar.

The Tok Pisin Synthesizer

Piggy-backing the Swedish Synthesis

- In order to produce full speech-to-speech translation, a Tok Pisin synthesizer was needed.
- A makeshift Tok Pisin synthesis was obtained by piggy-backing on the already existing Swedish synthesizer [6].
- The only thing that was needed was a Tok Pisin lexicon, using Swedish transcriptions. An excerpt from the lexicon is shown below:

```
i [s] /'i/
kirap [s] /ki:'rap/
laik [s] /'lajk/
lo [s] /'På/
long [s] /'Pån/
lusim [s] /'Tos:im/
```

- The following five sentences were synthesized and played to native speakers on location on New Ireland. (SOUND FILE ON TAPE)
- Informal testing showed that whereas sentence (1) was very hard to understand, sentences (2) to (5) were understandable, despite strong Swedish accent.

Synthesized Sentences

(1) Mi laik painim ron bilong balus i kirap long Fraide.
(I would like a flight that departs on Friday.)

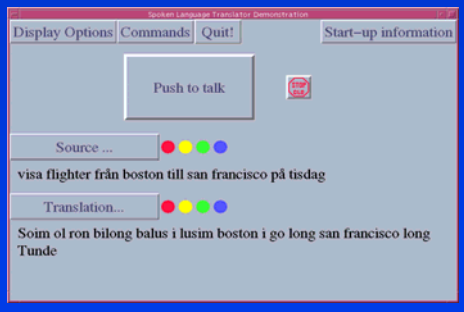
(2) Hamas bilong baim tiket i go long Mosbi?
(How much does a ticket to Port Moresby cost?)

(3) I gat sampela sit i stap long dispela ron bilong balus?
(Are there any seats left on that flight?)

(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.)

(5) Wanem taim bai mi lusim dispela hap?
(What time do I leave?)

Translation: SLT Demonstrator Interface



Demonstration example: Swedish input sentence and Tok Pisin translation. (SOUND FILE ON TAPE)
English translation: Show flights from Boston to San Francisco on Tuesday.

Summary

Evaluation and Future Work

- Although all links in the translation chain has a rudimentary or makeshift character, a running system has been created.
- Due to the minimality of the system, a formal evaluation of the system was not attempted. Moreover, there was also the problem of finding native speakers of Tok Pisin in Sweden.
- An ATIS corpus in Tok Pisin has been created, which despite its small size, provides information concerning a small, but interesting, number of idiomatic expressions.
- The language module so far only covers a small set of word-to-word rules. An implementation of a real Tok Pisin grammar and transfer rules at QLF level is yet to be done.
- The synthesizer has proven to be of some usability, but its Swedish heritage makes it sub-optimal. Much improvement could probably be obtained by writing Tok Pisin-specific prosody rules.
- The ease with which a running system was created shows how well a modular system such as the SLT can be expanded to encompass new language pairs.

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 (Kalämendy) 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.

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, F., Wong, P. & 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.