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”, excluding 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 kanz*) 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 *pins*. 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.
**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 unidirectional word-to-word transfer rules have been implemented. An example of such a rule looks like this:

```
rule_wv([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 New Ireland.

---

**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.

---

**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 obtained by writing Tok Pisin-specific prosody rules for a 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 Wiren; 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


