A grammatical analyser for Tokelau

Trond Trosterud
UiT The Arctic University of Norway
trond.trosterud@uit.no

Arnfinn Muruvik Vonen
Oslo Metropolitan University
arnfinn.vonen@oslomet.no

Abstract

This article will present a grammatical analyser, disambiguator and dependency analysis of Tokelau. The grammatical analyser is written as a finite-state transducer (FST), whereas the disambiguator and dependency analyser are written in Constraint Grammar (CG), both within the GiellaLT infrastructure. Contrary to most languages analyzed within this framework, Tokelau is a Polynesian language and thus predominantly isolating language, with reduplication and affixation as the main morphological processes. The discussion on Tokelau will thus also be relevant for an FST and CG treatment of other Polynesian languages.

1 Introduction

This article will present a lexicon, morphological analyser, disambiguator and dependency grammar for Tokelau.

Section 2.1 gives a background of the Tokelau language and the grammatical approach behind the analysis. Section 3 presents the morphological analysis, section 4 discusses disambiguating the Tokelau morphology and analysing it syntactically via a dependency analysis. Section 5 contains an evaluation of the grammatical models, section 6 discusses practical tools derived from the analysers. Finally comes a conclusion.

2 Background

2.1 The Tokelau language

Tokelau, also known in English as Tokelauan, is a Polynesian language, spoken on the Tokelau Islands, a dependent territory of New Zealand located between Samoa and Kiribati in the Pacific Ocean. Being a Polynesian language, Tokelau has a mainly isolating word structure. Affixation is used for such purposes as causativization and nominalization. One interesting category with respect to grammatical analysis is verbal number, which is expressed morphologically on the verb in several ways, particularly reduplication and prefixation, depending partly on the shape of the verbal stem. Verbal number, if present, agrees with the absolutive (bare) noun phrase (see 4.4 below on Tokelau sentence structure). Oblique arguments and adverbials are preposition phrases (Simona, 1986a, p. xxix).

There are comprehensive studies of Tokelau grammar, notably a dissertation on Tokelau noun phrase structure (Vonen, 1997) as well as one on Tokelau syntax in general (Hooper, 1993). The standard dictionary, Simona (1986b), also contains a lengthy grammatical sketch (Simona, 1986a). There are two grammatical handbooks on the language (Hovdhaugen et al., 1989; Hooper, 1996).

The main non-segmental morphological process is syllabic reduplication, where the first syllable of the stem is reduplicated, as shown in Simona (1986a) p. xxvi-xxvii. Example (1) shows the singular and plural forms of two verbs.

(1) a. nofo - nonofo ‘sit, live (sg-pl)´
b. galue - gālulue ‘work’

There are also examples of total reduplication, although only on bisyllabic stems. Example (2) shows a shift from neutral to continuative Aktionsart:

(2) a. alo – aloalo
‘paddle – paddle continuously’
b. logo – logologo
‘tell – tell everyone’

Reduplication has several functions in Tokelau, expressing plural is only one of them. For lexicalised reduplication the best way will be to just list reduplicated forms in the lexicon. Working on the present analyser will help distinguishing between productive and lexicalised reduplication.

Nouns are not inflected. Number is indicated by means of prenominal determiners denoting either singular or plural.

2.2 The grammatical framework behind this study

This study presents a grammatical model of the Tokelau language, using Finite State Transducers and Constraint Grammar, as presented in sections 3 and 4 below. The code itself uses the *GiellaLT* infrastructure (giellalt.github.io). This infrastructure consists of a language-independent set-up of build routines for turning grammatical models into programs analysing running text as well as into practical programs like spellcheckers, grammarcheckers, etc.

A more thorough presentation is given in Moshagen et al. (2023).

2.3 Earlier research on Polynesian language models

To our knowledge, there have not been any attempt so far at building language models for Tokelau[1]. A related work for Māori is Finn et al. (2022b). The main point for the authors is to established a tagset for Māori based upon Māori grammar instead of just copying the tagset from the Universal Dependency framework[2], referred to as a tagset "not fit for non-European languages" (op.cit. p. 6). The same authors also investigate the role the analysis of particles plays in Māori disambiguation, cf. Finn et al. (2022a). Their conclusion is that the part-of-speech (POS) label PARTICLE is glossing over grammatical differences within the class. As a case in point they quote the Māori particle *pai* 'good, well', capable of (pre)modifying both verbs and nouns, thus calling for the terms *adverb* and *adjective*. The authors argue against this, and argue that since *pai* in both cases modifies the word it preceeds, be it nouns or verbs, a POS label MODIFIER (MOD) is a better option. We will return to this issue, arguing for a third solution, in sections 4.1 and 4.3.

Karnes et al. (2023) presents an analysis of Māori based on Universal Dependency. The paper is relevant in this context, since we, too, will present a dependency analysis of Tokelau. Our approach will differ from their UD approach in some respects: In UD, lexical words are mothers of functional words, whereas in our approach functional words act as mothers to lexical words when they govern the distribution of the functional words. Thus, in what phrase structure grammars analyse as PPs, UD sees the preposition as the daughter of the noun, whereas in our (Constraint Grammar) approach the noun is the daughter of the preposition, thus making it possible to distinguish between different verbal arguments. For (phrase structure) PPs, the P is the daughter of the verb, whereas for NPs, the N is the daughter. Within UD, this distinction must be made indirectly, distinguishing between N daughters *without* P daughters (objects etc.) and N daughters *with* P daughters (PPs in phrase structure frameworks). An analysis within our framework may still be converted to UD and vice versa.

3 The Tokelau Finite State Transducer

The Tokelau grammatical model is written as a finite state transducer, as presented in Beesley and Karttunen (2003). The morphophonological rules were written in *twolc*, as first presented in Koskenniemi (1983).

3.1 Morphophonology

With relatively little morphology, the main challenge for the Tokelau morphophonological component is reduplication.

Partial reduplication in finite state transducers was solved by Beesley and Karttunen (2003) (p. 487-493), for twolc. The idea is to use the reduplicating CV sequence as the reference point, and build one rule to copy the C and another to copy the V. In lexc, the reduplication morphology is represented as `^R^E>`, where the > marks the stem boundary and the `^R` and the `^E` the position to copy the C and V, respectively. This is then pointed to the stem lexicon, where we find the reduplicating stems, e.g. *nofo* 'sit'. This string is then applied to the twolc reduplication rules:

```
^R:Cx <=> .#. _  ^E: %> Cx ;
    where Cx in Cns ;

^E:Vx <=> .#. ^R: _  %> (Cns) Vx ;
    where Vx in Vow ;
```

[1] The single exception, less relevant in this context, is Kargaran et al. (2023), which contains language identification models for 1665 languages, one of them for Tokelau.

[2] The authors do not refer to any version of the Universal Dependencies (UD) tagset, but the tagset of UD Version 2 may be found at universaldependencies.org/u/pos/.

In the `R:Cx` rule, `^R` gets it value from the variable `Cx` (for this stem, n, and in the corresponding rule `E:Vx`, `^E` gets it variable from `Vx`, here e. The result is the reduplicated form $nonofo$ 'sit (plural)'.

For full reduplication, Beesley and Karttunen (2003) enclosed the string to be reduplicated with two boundary symbols ^[and ^], and then denoted the duplication of the string with an operator, written ^2. An algorithm **compile-replace** would then duplicate the string. For their example language Malay, this would then result in plural forms like *bukubuku* and *pelabuhanpelabuhan* from *buku* 'book' and *pelabuhan* 'port', respectively.

Now, this algorithm is not available in *hfst-lexc*, the compiler used for the present model of Tokelau. What we instead did, was to utilise the fact that the so-called full reduplication of Tokelau in practice only involves two syllables. We thus extended the analysis above to two syllables. For two-syllable reduplication, we made 4 rules, one for each sound. The two last ones, closest to the stem, were as the ones above, and the two first ones were as follows:

```
^R1:Cx <=> _ ^E1: ^R2: ^E2: %> Cx ;
     where Cx in Cns ;

^E1:Vx <=> _ ^R2: ^E2: %> (Cns) Vx ;
       where Vx in Vow ;
```

Here, each of the four reduplicated letters gets its own value, and the result is *logologo* from `^R1^E1^R2^E2>` pointing to *logo*. Note that the same rule will also work for *aloalo*, only with reference to the last three letters.

Since there is only a finite number of verb stems marking plural via reduplication, an alternative option would have been to have two entries in the lexicon, one for the singular form and one for the plural form, both with the singular form as lemma. In our experience, lexicon maintenance is easier with one entry per lemma. A further benefit of modeling the morphological processes explicitly is that it gives a transparent picture of the morphological processes of the language in question, here Tokelau.

3.2 Lexicon and morphology

Tokelau has two open parts of speech, nouns and verbs. For the nouns, we identified a subgroup *location nouns*, the rest were simply given a `+N` tag.

For the verbs, we modeled number as a morphological process. The reduplication prefixes were added prior to the stem, as shown in section 3.1 above. Some verbs form their plural not with reduplication but with adding a prefix *ta-*, for these we simply added the prefix. Some verbs do not form a distinct plural form, they were marked as ambiguous.

The words belonging to the closed classes were added and provided with tags reflecting their syntactic behaviour.

4 The Tokelau Constraint Grammar

The constraint grammar framework was originally presented in Karlsson (1990). The present implementation is written in *vislcg3*, as presented in Bick (2023).

To put it simply, we distinguished 4 types of particles and based the N/V disambiguation as well as the NP delimitation on four types of functionally defined particles, defined by whether they precede or follow their head word, and by whether the head word is a verb or a noun[3]. Thus, contrary to previous research, we argue for more rather than less types of particles, dependent upon the function they play in the sentence, and in our perspective, dependent upon their ability to delimit noun phrases and verbs.

4.1 Noun phrases

The structure of noun phrases that we are adhering to, is largely in line with the analyses found in Hooper (1996) and Hovdhaugen et al. (1989), although we disagree with them in viewing the preposition not as a part of the noun phrase, but rather as forming a preposition phrase with the NP.

The noun phrase, in our analysis, consists of a determiner (which may, in certain cases, be null), a possible premodifier (typically, a size indicator), a nucleus (a common noun, a proper noun, a locative noun, or a personal pronoun), and possibly one or more postmodifiers (lexical modifiers and/or postmodifying particles) (Example from Simona (1986a, xix).

(3) he mātuā ika
this.DET.INDEF:SG big.PCLE.PRE fish.N.SG
lele
very.PCLE.POST
'a huge fish'

[3] This is a slight simplification. We also distinguish some specific particles modifying proper nouns or numerals. We will not discuss them here.

The constraint grammar assigns the tag @>N to prenominal determiners and particles and the tag @N< to postnominal demonstratives and particles. Dependency rules then set a dependency relation between the noun and its modifiers, as long as no non-NP constituents (NPNH) occur between the determiner or demonstrative and the noun.

```
MAP (@>N) TARGET Det IF
    (*1 N BARRIER NOT-PRE-N);
MAP (@>N) TARGET Pre IF
    (*1 N BARRIER NOT-PRE-N);
MAP (@N<) TARGET Post IF (*-1 N BARRIER V);
MAP (@N<) TARGET Dem IF (-1 N)(NOT 1 N);

SETPARENT @>N TO (*1 N BARRIER NPNH);
SETPARENT @N< TO (*-1 N BARRIER NPNH);
```

4.2 Preposition phrases

Most noun phrases are complements of prepositions. In the Tokelau constraint grammar, the PP structure is expressed by two rules. First, a MAP rule assigns the syntactic tag @P< ("I am a complement to a **P** to my left") to the N closest to the P. Then, a SETPARENT rule sets a dependency relation between the preposition and the noun complement.

```
MAP (@P<) TARGET N OR Pron IF
    (*-1 Pr BARRIER N OR V);
SETPARENT:r6 @P< TO (*-1 Pr BARRIER V);
```

4.3 The verbal complex

The term *verb phrase* (VP) is usually understood as "the verb and its arguments (except the subject)". Tokelau syntax does not quite work this way, and we will thus avoid the term VP for Tokelau. Instead, we identify the *verbal complex*, by which we mean the main verb, its auxiliaries and verbal particles. Being an isolating language, Tokelau expresses both morphosyntactic categories related to verbs (tense-aspect, negatives) and Aktionsart as verbal particles.

The verbal complex, too, is analysed largely according to Hooper (1993, 52f), but see also Hooper (1996) and Hovdhaugen et al. (1989). The verbal complex consists of a tense-aspect particle (absent in certain functions such as the imperative or the narrative relating of series of events), a possible negative particle, a possible preverbal subject pronoun, a possible prenuclear particle or auxiliary, then the verb. After the verb may follow a postmodifying particle, a possible directional particle, a possible anaphoric particle, and a possible manner particle (Hooper, 1993).

In addition to their occurrence as predicates in verbal sentences, verbs may also occur in nominalized structures, expressed by preposing the singular definite article *te*. In these constructions, tense-aspect particles usually do not occur. The absolutive phrase is replaced with a possessive phrase. These nominalized constructions may be expanded by a nominalizing suffix *-ga*, usually to express past tense. The position of the suffix is usually after the directional particles, and sometimes additionally before the directional particle in (4) (examples, glosses and translations from p. 35 in Hooper (1996):

(4) a. Kāmata loa toku havalivali
begin MAN 1SG.POSS walk.REDUP
mai ki te kakai...
DIR to.PR DET village
'My [habit of] walking to the village began then...'

b. Ko te galo atuga lava tēnā
PR DET disappear DIR-NOM INT DEM
o Lata
of L.
'That was the complete disappearance of Lata.'

The verbal complex is analysed in the same way as the noun phrases discussed in section 4.1.

4.4 Sentence structure

Tokelau is generally considered a verb-initial language. That is, the verbal complex often introduces the sentence, and if a topical or focused noun phrase is preposed to the verbal complex, they are usually marked with the "presentative" preposition *ko*.

The case-marking pattern of the language is ergative: An NP referring to the main argument of an intransitive verb (cf. "non-agentive sentence" in Hovdhaugen et al. (1989) and Hooper (1996)) or the patient of a transitive verb (cf. "agentive sentence" in Hovdhaugen et al. (1989) and Hooper (1996)) is marked as absolutive (i.e., without a preposition), while an NP referring to the agent of a transitive verb is marked as ergative (called "agentive" in Hovdhaugen et al. (1989) and Hooper (1996)) with the preposition *e*. Also verbal number, to the extent that it expresses agreement with the number of an argument of the verb, follows the ergative pattern.

Being an ergative language, Tokelau marks the single argument of an intransitive verb (called **S**)

Test	Words	Coverage
First test	351080	0.9026
3 weeks' development	320540	0.9656

Table 1: Coverage of the New Testament

in the same way as the patient argument of a transitive verb (called **P**), whereas the agent argument is marked differently, with the preposition *e* (called **A**). In Constraint Grammar, both the S and the P arguments will be identified as NPS linked to the verb without any intervening preposition for the P also with a PP headed by an *e* preposition to its right or left.

In addition to verbal sentences, the language allows locative, possessive and nominal sentences. Locative and possessive sentences resemble verbal sentences, but they have a locative (*i* 'in, at') or possessive (*a*, *o*) prepositional phrase in the verbal slot in the verbal complex. Nominal sentences include a presentative prepositional phrase in the verbal slot and usually have no tense-aspect marker or other pre- or post-modifiers characteristic of the verbal complex.

If the NP preceded by the ergative preposition *e* is a personal pronoun, then this referent may alternatively or additionally be expressed by a preverbal pronoun. A preverbal pronoun may also be used if the verb is intransitive, but only if the verbal complex contains the dehortative tense-aspect marker *nahe*.

5 Evaluation

5.1 The FST

The grammatical model was developed by working with the Tokelau grammars and dictionary (Simona, 1986b) and testing it successively against the Tokelau New Testament, including adding names from the New Testament to the grammar model.

The lexicon contained 5780 lemmas during the final test, slightly less during the first one. An important result to notice is that even with a minimal morphology and a very small set of lemmas we were quickly able to achieve a coverage above 95 %.

Testing for text not used for development, we took a totally new genre, the Level 4 books from the Tokelau Ministry of Education (gagana-tokelau.org.nz), containing books on a wide range of topics, solar energy, COVID, coral bleaching, ancient navigation, poetry and local governance, to mention a few. The books contained many new names that, contrary to the New Testament, had not been added to the language model. We thus tested the corpus twice for coverage, with and without words having an initial capital letter. Table 1 shows the result for the New Testament and table 2 shows the results for the book corpus. In table 1, "First test" refers to testing the coverage based upon the general lexicon only whereas "3 weeks' development" refers to results after having improved the grammar and (above all) added missing names and words to the lexicon. Needless to say, improving the coverage beyond 96.56 % is certainly doable, the lexicon just needs more work.

Test	Words	Coverage
The full text	62281	0.8897
Excl. capitalised words	44442	0.8976

Table 2: Coverage: Min. of Education books

Type	Analyses	An / words
Without disamb	637469	1.66
With disamb	383837	1.17

Table 3: Disambiguating the New Testament

Also for table 2, the coverage is quite good, almost 89 % for the full text and almost 90 % when excluding names.

5.2 The constraint grammar

We tested the constraint grammar on the New Testament.

Table 3 shows that each Tokelau word has on average 1.66 different analyses. 21 constraint grammar rules have reduced this number to 1.17, and increasing the number of constraint grammar rules will no doubt bring the disambiguation ratio closer to 1.0.

6 Practical tools

6.1 The spellchecker

The GiellaLT infrastructure offers a ready-made setup for converting transducers into spellcheckers. We did that for Tokelau[4].

A controversial issue in Tokelau orthography is the representation of vowel length, which is

[4] The resulting program may be downloaded via the *Divvun Manager*, available at divvun.no and put to use in Windows, Macintosh and Linux computers.

Test	Words	1st pos	Top 5
NT, long V	315	95.9	98.7

Table 4: Correcting missing length mark

phonemic and may be indicated by a macron above a vowel to indicate that it is long. According to Ministry for Pacific Peoples (2024, p. 7), ”support for and against consistent macron use is broadly divided along diaspora and Tokelau lines respectively”. Thus, the Government of Tokelau (Matāeke o Akoakoga a Tokelau, n.d.) provides guidelines such as ”Don’t include macrons where the pronunciation is widely known”, while the New Zealand Ministry of Education (Ministry of Education, 2009, p. 14) and Ministry of Pacific Peoples (Ministry for Pacific Peoples, 2024) emphasize the usefulness of macrons for the language learner. In the texts we have worked on so far, there is a tendency to drop the macron whenever a corresponding word with a short vowel does not exist. 55.5 % of the words with long vowel in the Tokelau New Testament recognised by the language model did not have a short vowel counterpart, thus adding vowel length distinguished minimal pairs in 44.5 % of the cases.

In the further development of the spellchecker, we will approach the relevant authorities to make sure that the spellchecker will be in line with the needs of the Tokelau language community. A spellchecker able to correct macron errors (missing or hypercorrect long vowel marks) may easily be changed into a spellchecker tolerating missing macrons or even into one prohibiting macrons.

In order to test the spellchecker’s ability to correct missing length marks, we extracted all words containing one long vowel from the New Testament, and removed the length mark. We then made sure that the resulting word was unknown to the language model. This resulted in a t of 315 distinct wordforms. We ran this list through the spellchecker, and measured whether it was able to suggest the correct form as the first correction suggestion or as one of the five corrections in the top five-list (since most spellcheckers offer only 5 suggestions, suggestions further down the list of suggestions were ignored). It turned out that the spellchecker was able to correct length mark omission on 98.7 % of the cases, in 95.9 % of the cases the correction was given in the first position on the suggestion list, cf. table 4.

The cases where the targeted form was not the first suggestion were either forms with more than one potential length error or forms where other suggestions were common words (cf. (5)).

(5) **Halamo* → *Halāmo, Halamō*
**Ha* → *Ma, la, Na, Ka, La, Hā*

6.2 The need for a grammarchecker

Since a large part of Tokelau word tokens are short (one and two letters long), we predict the number of real-word errors to be larger than for synthetic languages, having longer words.

Our hypothesis is thus that in an isolating language like Tokelau, a smaller part of text correction is actually linked to orthographic errors, and a larger part to grammatical errors, to what to the proofing tool will look like wrong use of shorter words.

Building a grammarchecker falls outside the scope of the present paper and is left for future research.

7 Conclusion

We have during a short time built a finite state transducer, constraint grammar disambiguator, syntactic function and dependency relation annotator for Tokelau, a Polynesian language spoken on the Tokelau islands and in diaspora communities in New Zealand and elsewhere. The empirical basis for the programs was the standard Tokelau dictionary ((Simona, 1986b)), as well as the standard grammars for the language ((Hooper, 1993, 1996; Vonen, 1997; Hovdhaugen et al., 1989). The result was a language model with relatively high coverage (89 % on unseen text, excluding proper nouns) and relatively high disambiguation rate (1.17 readings/wordform). The resulting transducer has been implemented as a spellchecker for Microsoft Word for Windows and for standard Macintosh programs. Testing shows that it may correct the most common spelling error in Tokelau (vowel length) quite efficiently, 98.7 % of a set of artificially created length errors were corrected, 95.9 % of them as the first suggestion.

Acknowledgments

Thanks to Sjur N. Moshagen for helping out with the implementation of the Tokelau keyboard and proofing tools.

References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. Studies in Computational Linguistics. CSLI Publications, Stanford, California.

Eckhard Bick. 2023. Visl & cg-3: Constraint grammar on the move: An application-driven paradigm. In *Rule-Based Language Technology*, volume 2 of *NEALT Monograph Series*, pages 112–140, University of Tartu. NEALT.

Aoife Finn, Suzanne Duncan, Peter-Lucas Jones, Gianna Leoni, and Keoni Mahelona. 2022a. Annotating "particles" in multiword expressions in te reo Māori for a part-of-speech tagger. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 67–74, Marseille, France. European Language Resources Association.

Aoife Finn, Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2022b. Developing a part-of-speech tagger for te reo Māori. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 93–98, Dublin, Ireland. Association for Computational Linguistics.

Robin Eliszabeth Hooper. 1996. *Tokelauan*, volume 58 of *Languages of the world Materials*. Lincom Europa, München - Newcastle.

Robin Elizabeth Hooper. 1993. *Studies in Tokelauan syntax*. University of Auckland, Auckland.

Even Hovdhaugen, Ingjerd Hoëm, Consulata Mahina Iosefo, and Arnfinn Muruvik Vonen. 1989. *A handbook of the Tokelau language*. Norwegian University Press and The Institute for Comparative Research in Human Culture, Oslo.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLING '90 Proceedings of the 13th conference on Computational linguistics*, volume 3, pages 168–173, Helsinki.

Sarah Karnes, Rolando Coto, and Sally Akevai Nicholas. 2023. Towards Universal Dependencies in Cook Islands Māori. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 124–129, Remote. Association for Computational Linguistics.

Kimmo Koskenniemi. 1983. *Two-level Morphology. A General Computational Model for Word-forms Production and Generation*, volume 11 of *Publications of the Department of General Linguistics*. University of Helsinki.

Sjur Nørstebø Moshagen, Flammie Pirinen, Lene Antonsen, Børre Gaup, Inga Mikkelsen, Trond Trosterud, Linda Wiechetek, and Katri Hiovain-Asikainen. 2023. *The GiellaLT infrastructure: A multilingual infrastructure for rule-based NLP*, volume 2 of *NEALT Monograph Series*, pages 70–94. NEALT.

Romati Simona. 1986a. An outline of Tokelau grammar. In *Tokelau Dictionary*, pages xi–xlix. Office of Tokelau Affairs, Western Samoa.

Romati Simona. 1986b. *Tokelau Dictionary*. Office of Tokelau Affairs, Western Samoa.

Arnfinn Muruvik Vonen. 1997. *Parts of speech and lingustic typology. Open classes and conversion in Russian and Tokelau*, volume 22 of *Acta Humaniora*. Det historisk-filosofiske fakultet , University of Oslo.