Tense and Aspect in Runyankore using a Context-Free Grammar

Joan Byamugisha and **C. Maria Keet** and **Brian DeRenzi**
Department of Computer Science, University of Cape Town, South Africa,
{jbyamugisha,mkeet,bderenzi}@cs.uct.ac.za

Abstract

The provision of personalized patient information has been encouraged as a means of complementing information provided during patient-doctor consultations, and linked to better health outcomes through patient compliance with prescribed treatments. The generation of such texts as a controlled fragment of Runyankore, a Bantu language indigenous to Uganda, requires the appropriate tense and aspect, as well as a method for verb conjugation. We present how an analysis of corpora of explanations of prescribed medications was used to identify the simple present tense and progressive aspect as appropriate for our selected domain. A CFG is defined to conjugate and generate the correct form of the verb.

1 Introduction

In Uganda, patients receive medical information verbally during the patient-doctor consultation. However, DiMarco et al., (2005; 2006) and Wilcox et al., (2011) noted that patients consistently retain a rather small fraction of the verbal information after the consultation, possibly resulting in improper compliance to medical instructions. Further, it was found that personalized information increases the likelihood for a patient to be more engaged and likely to read, comprehend, and act upon such information better (Cawsey et al., 2000; Wilcox et al., 2011).

The fundamental complexity in the customization of patient information is the number of different combinations of characteristics, which can easily be in the tens or hundreds of thousands (DiMarco et al., 2005). Natural Language Generation (NLG) has successfully been applied to generate personalized patient information (DiMarco et al., 2005; DiMarco et al., 2006; De Carolis et al., 1996; de Rosis and Grasso, 2000).

Localized patient information is encouraged because the use of English exacerbates literacy difficulties already prevalent in situations of health (DiMarco et al., 2009). Our broader programme of NLG for Bantu languages aims to apply NLG to generate drug explanations in Runyankore—a Bantu language indigenous to Uganda, where English is the official language, but indigenous languages are predominantly spoken in rural areas. Runyankore sentences generated through ontology verbalization (Byamugisha et al., 2016) exposed two crucial issues: (1) *What tense and aspect is used in explanations of prescribed medication?* and (2) *Is a context-free grammar (CFG) sufficient to conjugate verbs in Runyankore?* Through the analysis of two relevant corpora, we identify that the simple present (universal) tense with the progressive aspect would be best for generating explanations of prescribed medications. We demonstrate that this can be done for Runyankore using a CFG for verb conjugation.

In the rest of the paper, we first summarize the Runyankore verbal morphology (Section 2) and related work (Section 3). Section 4 presents the corpus analysis. The relevant CFGs for the Runyankore verb are presented in Section 5. We discuss in Section 6 and conclude in Section 7.

2 Verbal Morphology of Runyankore

Runyankore is a Bantu language spoken in the south western part of Uganda by over two million people, which makes it one of the top five most populous languages in Uganda (Asiimwe, 2014; Tayebwa, 2014; Turamyomwe, 2011). Like other Bantu languages, it is highly agglutinative to the extent that a word can be composed of over five constituents (Asiimwe, 2014; Tayebwa, 2014). Runyankore has twenty noun classes, (NC), and each noun belongs to a specific class.

Our discussion of tense and aspect in Runyankore throughout this paper is based on work done by Turamyomwe (2011). The standard classification of tense by dividing time into past, present, and future is further subdivided, resulting in fourteen tenses. Aspect focuses on the internal nature of events, instead of their grounding in time. There are two major aspects: the perfective and imperfective the latter subdivided into persistive, habitual, and continuous, with the progressive as a subtype of the continuous). Runyankore expresses tense as prefixes and aspect as affixes to the right of the verb stem. Table 2 shows the different 'slots' in Runyankore's verbal morphology. We illustrate the general structure of the Runyankore verbal morphology, where neg is negation, RM is remote past, VS is verb stem, App is applicative, FV is final vowel, Loc is locative, Emp is emphatic, and Dec is declarative.

- `titukakimureeterahoganu`
  'We have never ever brought it to him'
- `ti tu ka ki mu reet er a ho ga nu`
- neg-(NC2 SC)-RM-(NC7 SC)-(NC1 SC)-VS-App-FV-Loc-Emp-Dec

The compulsory slots are the *initial*, *formative* (except in the case of the universal and near past tense), *verb-stem*, and *final*.

Slot	Grammatical Category	Morpheme
pre-initial	1. primary negative 2. cont. marker	1. ti- 2. ni-
initial	subject marker	depends on the NC
post-initial	secondary negative	-ta-
formative	tense	all tenses except near past
limitative	persistive aspect	-ki-
infix	object marker	depends on NC
extensions	App; Cs; Ps; Rec; Rev; Stv; Itv; Red; Ism	-er-, -erer-, -ir-; zi-, -is-; -w-; -n-; -ur-, -uur-; -gur-; repeat the stem; -is+ pre-initial
final	1. final vowel (a) indicative, (b) subjunctive 2. near past tense	1. (a) -a (b) -e 2. -ire
post-final	1. locatives 2. emphatic 3. declarative	1. -ho, -mu-yo 2. -ga 3. -nu

Table 1: Verbal Morphology of Runyankore (Turamyomwe, 2011); App: applicative, Cs: causative, Ps: passive, Rec: reciprocal, Rev: reversive, Stv: stative, Itv: intensive, Red: reduplicative, Ism: instrumental

3 Related work

We center our discussion here around the existing methods of verb conjugation for tense and aspect in agglutinated languages like Tamil (Rajan et al., 2014) and Turkish (Fokkens et al., 2009). The placement of morphemes in a word, and rules governing the combinations of morphemes to form semantic categories are important in agglutinated languages (Jayan and Bhadran, 2015). Similar to Runyankore, the sequence of morphemes can express mood, tense, and aspect (Rajan et al., 2014; Fokkens et al., 2009; Turamyomwe, 2011).

There are several approaches for text generation in agglutinated languages, being corpus-based, paradigm-based, Finite-State Transducer (FST)-based, rule-based, and algorithm-based (Antony, 2012). Some of these are currently inapplicable to Runyankore because it is structurally different or too under-resourced. We thus decided to implement tense and aspect in Runyankore using a rule-based approach, derived from a set of grammar rules and a dictionary of roots and morphemes. A CFG is powerful enough to depict complex relations among words in a sentence, yet computationally tractable enough to enable efficient algorithms to be developed (Jurafsky and Martin, 2007). Because the verb conjugation work presented here is intended to be one of the components in a Runyankore grammar engine, the use of a CFG is justified.

4 Tense and Aspect in Prescription Explanations

To the best of our knowledge, there is no prior work specifically discussing tense and aspect for explanations of prescribed medications. We instead analyze text describing drug prescriptions from empirical studies (Berry et al., 1995; Berry et al., 1997).

We limited our analysis here to the corpora from Berry et al., (1995; 1997), compiled from a series of empirical studies done to ascertain the kind of information patients and doctors considered important about prescribed medication. We further only considered the tense in the main clause of the sentences in the corpus in order to simplify our initial scope.

We analyzed 27 sentences, 18 from (Berry et al., 1997) and 9 from (Berry et al., 1995), describing medication prescriptions. We were interested in the form of the verb, in order to identify the tense and aspect used. Table 2 shows how often each verb form occurred in each unique sentence in the corpus.

Example	**Tense, Aspect**	**|Occ.|**
have	simple pres. ind.	2
reduce	simple pres. ind.	1
is	simple pres. ind.	5
should take	pres. imp.	3
contains	simple pres. ind.	1
are	simple pres. ind.	3
if it does not relieve	pres. cond.	3
may be taken	past perf. subj.	1
may cause	simple pres. subj.	2
should be avoided	past perf. imp.	2
do not contain	simple pres. ind.	1
to store	infinitive	1
are produced	pres. perf. ind.	2

Table 2: Tense and Aspect used in Prescription Explanations; pres.=present, perf.=perfect, ind.=indicative, subj=subjunctive, imp.=imperative, cond.=conditional, occ.=occurence

The simple present tense is used in 55.5% of the corpus, in 48.2% with the indicative aspect, and in 7.7% with the subjunctive. The simple present tense and indicative aspect is used in those sentences which are informational in nature, but the present tense and imperative aspect for those which are instructional (for example 'should take Fennodil ...' and 'should adopt a more suitable ...').

5 Verb Conjugation using a CFG

We devise a CFG for verb conjugation in the simple present tense (Runyankore's 'universal' tense), and the auxiliary 'has' and copulative 'is' (from 'to be') as special cases that do not conform to the standard grammatical structure.

5.1 Universal Tense in Runyankore

The universal tense has no special tense marker, and as such is sometimes called the null tense (Turamyomwe, 2011). We apply the progressive aspect, which marks a situation which is ongoing at the time of use. This is appropriate for informational sentences such as those listed in Section 4, because this information will always be true as long as one is on that medication. We introduce a new non-terminal, *initial group*, which, depending on the tense and aspect applied, has productions for one or more of the three 'initial' slots (cf. Table 2). We only consider five slots here: the *pre-initial*, as well as the four compulsory 'slots' discussed in Section 2. We assign all six nonterminals the symbols: IG for initial group, PN for pre-initial, IT for initial, FM for formative, VS for verb-stem, and FV for final vowl. Finally, since this tense has no tense morpheme, we will use the production $FM \rightarrow \emptyset$ to illustrate it. The example shows productions with verb stems *kyendez* 'reduce,' *gw* 'fall,' *vug* 'drive,' and *gend* 'go':

$S \rightarrow IG\ FM\ VS\ FV$
$IG \rightarrow PN\ IT$
$PN \rightarrow$ ti | ni
$IT \rightarrow$ a | o | n | tu | mu | ba | gu | gi | ri | ga | ki |
bi | e | zi | ru | tu | ka | bu | ku | gu | ga
$FM \rightarrow \emptyset$
$VS \rightarrow$ kyendez | gw | vug | gend
$FV \rightarrow$ a | e | ire

The production of IT has several possible values, depending on the noun class of the subject of the sentence. For all verbs, except 'has' and 'to be,' FV will always be the indicative final vowel 'a'.

5.2 Deviations from Standard Grammar

There are two verbs which deviate from the standard Runyankore grammar: the auxiliary 'has' (verb stem *in*) and the copulative 'to be' (verb stem *ri*). The auxiliary deviates in two main ways: first, the continuous marker is dropped, and second, the subjunctive final vowel 'e' is used instead. The copulative deviates even further because it both drops the pre-initial and has no final vowel. It is thus our design decision to use separate CFGs for these special cases, for two main reasons: firstly, to prevent the generation of sentences like *nibaina, niguine* or *nibaria, nigurie* which do not exist in the language. Secondly, there is no way to limit the inclusion of $\emptyset$ as a terminal for PN and FV to only these special cases, instead of having it applied to all verbs. The CFG for 'has' (verb-stem *-in-*):

$S \rightarrow IG\ FM\ VS\ FV$
$IG \rightarrow PN\ IT$
$PN \rightarrow \emptyset$
$IT \rightarrow$ a | o | n | tu | mu | ba | gu | gi | ri | ga | ki | bi | e | zi | ru | tu | ka | bu | ku | gu | ga
$FM \rightarrow \emptyset$
$VS \rightarrow$ in
$FV \rightarrow$ e

The CFG for the case of 'to be' (verb-stem *-ri*) is almost the same as for 'have', except for the following two production rules:

$VS \rightarrow$ ri
$FV \rightarrow \emptyset$

The CFGs show that verb conjugation can be achieved following the grammar rules on the verbal morphology. We have limited our non-terminals to six, only those necessary to generate text in our selected tense and aspect. However, by including more of the grammatical categories presented in Table 2, it would be possible to create the rules to generate many more tenses and aspects.

The patterns for generating Runyankore sentences from ontologies required a method for verb conjugation in order to generate correct text. We thus illustrate the use of the CFG in this context, using a sentence taken from the corpus by (Berry et al., 1997), which we modify and represent as a side effect in the example $Fennodil \sqsubseteq \exists hasSideEffect.Diarrhea$: *Buri Fennodil eine hakiri ekirikurugamu kitagyendereirwe kimwe ekya okwirukana* 'Each Fennodil has at least one side effect of diarrhea'. According to Byamugisha et al., (2016), *Buri* is the translation of 'each' for subsumption ($\sqsubseteq$), *eine* for 'has', *hakiri* for 'at least', and *kimwe* for 'one'; *ekirikurugamu kitagyendereirwe* is 'side effect', and *okwirukana* is 'diarrhea'. The *eine* 'has' has *e* as the subject prefix because Fennodil is placed in NC 9. With the CFG, one can thus generate several variations for 'has' that occur whenever a noun in a different NC is to the left of $\sqsubseteq$ in the axiom; for example *aine*, *baine*, *giine*, *riine* for NC 1, 2, 4, and 5 respectively.

6 Discussion

The identification of the tense and aspect relevant to our domain of interest—explanations of prescribed medications—through the analysis of corpora on medicine prescription enabled us to narrow down the scope of the text to be generated, in terms of tense and aspect, to only the simple present (universal) tense and continuous aspect. It is interesting that the present tense is appropriate for our target domain, because an ontology will be the input of our NLG system. Therefore, the consideration of generation of sentences, for example with the verb 'has,' mirrors axioms which either have 'has' as a role or the 'hasX' role naming, such as $hasSymptom$. In this way, our work here builds upon (Byamugisha et al., 2016) to verbalize ontologies in Runyankore, by solving two crucial issues: which tense and aspect to use, and how to achieve verb conjugation.

The use of CFGs allows for easy extensibility both to more tenses, and perhaps even other Bantu languages. For the case of tenses, we would only need to add new rules. The near past tense, for example, can be generated by changing the rule on FM from $FM \rightarrow \emptyset$ to $FM \rightarrow$ ka. CFGs for other Bantu languages can be produced by stating language-specific rules and terminals.

7 Conclusion

Through the analysis of corpora of prescription explanations, we identified that the simple present tense and progressive aspect were most suitable when generating informational drug explanations. Therefore, a CFG for universal tense, the auxiliary verb 'has', and the copulative was developed. Future work will include the implementation of these CFGs, inclusion of the imperative aspect, and evaluating the generated messages.

Acknowledgements

This work is based on the research supported by the Hasso Plattner Institute (HPI) Research School in CS4A at UCT and the National Research Foundation of South Africa (Grant Number 93397).

References

K P Antony, P J an Samon. 2012. Computational morphology and natural language parsing for indian languages: A literature survey. *International Journal of Scientific and Engineering Research*, 3.

Allen Asiimwe. 2014. *Definiteness and Specificity in Runyankore-Rukiga*. Ph.D. thesis, Stallenbosch University, Cape Town, South Africa.

C. Dianne Berry, Tony Gillie, and Simon Banbury. 1995. What do patients want to know: An empirical approach to explanation generation and validation. *Expert Systems with Applications*, 8:419 — 428.

C. Dianne Berry, C. Irene Michas, Tony Gillie, and Melanie Forster. 1997. What do patients want to know about their medicines and what do doctors want to tell them: A comparative study. *Psychology and Health*, 12:467–480.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2016. Bootstrapping a runyankore CNL from an isizulu CNL. In *5th Workshop on Controlled Natural Language*, Aberdeen, Scotland. Springer.

J. Alison Cawsey, B. Ray Jones, and Janne Pearson. 2000. The evaluation of a personalized health information system for patients with cancer. *User Modeling and User-Adapted Interaction*, 10(1):47–72.

Berardina De Carolis, Fiorella de Rosis, Floriana Grasso, Anna Rossiello, C. Dianne Berry, and Tony Gillie. 1996. Generating recipient-centered explanations about drug prescription. *Artificial Intelligence in Medicine*.

Fiorella de Rosis and Floriana Grasso. 2000. Affective natural language generation. In *Affective Interactions, LANI*, pages 204 – 218.

Chrysanne DiMarco, Peter Bray, Dominic Covvey, Don Cowan, Vic DiCiccio, Eduard Hovy, Joan Lipa, and Cathy Yang. 2005. Authoring and generation of tailored preoperative patient education materials. In *Workshop on Personalization in e-Health, User Modeling, Conference*, Edinburgh, Scotland.

Chrysanne DiMarco, Don Cowan, Peter Bray, Dominic Covvey, Vic DiCiccio, Eduard Hovy, Joan Lipa, and Doug Mulholland. 2006. A physician's authoring tool for generation of personalized health education in reconstructive surgery. In *American Association for Artificial Intelligence (AAAI) Spring Symposium on Argumentation for Consumers of Healthcare*, Stanford University.

Chrysanne DiMarco, David Wiljer, and Eduard Hovy. 2009. Self-managed access to personalized healthcare through automated generation of tailored health educational materials from electronic health records. In *American Association for Artificial Intelligence (AAAI) Fall Symposium on Virtual Health Interaction*, Washington D. C.

Antske Fokkens, Laurie Paulson, and M. Emily Bender. 2009. Inflectional morphology in turkish VP coordination. In *The HPSG09*, Germany.

V. Jayan and V. K. Bhadran. 2015. Difficulties in processing malayalam verbs for statistical machine translation. *International Journal of Artificial Intelligence and Applications (IJAIA)*, 6.

Daniel Jurafsky and H. James Martin. 2007. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Inc., USA.

K. Rajan, V. Ramalingam, and M. Ganesan. 2014. Machine learning of phonologically conditioned noun declensions for tamil morphological generators. *International Journal of Computer Engineering and Applications*, 4.

Doreen Daphine Tayebwa. 2014. Demonstrative determiners in runyankore-rukiga. Master's thesis, Norwegian University of Science and Technology, Norway.

Justus Turamyomwe. 2011. Tense and aspect in runyankore-rukiga: Linguistic resources and analysis. Master's thesis, Norwegian University of Science and Technology, Norway.

Lauren Wilcox, Dan Morris, Desney Tan, Justin Gatewood, and Eric Horvitz. 2011. Characterising patient-friendly micro-explanations of medical events. In *The SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*, pages 29–32, New York. ACM.