Natural Language Generation for Polysynthetic Languages: Language Teaching and Learning Software for Kanyen’kéha (Mohawk)

Greg Lessard
School of Computing
Queen’s University
Canada
lessard@cs.queensu.ca

Nathan Brinklow
Department of Languages,
Literatures and Cultures
Queen’s University
Canada
nathan.brinklow@queensu.ca

Michael Levison
School of Computing
Queen’s University
Canada
levison@cs.queensu.ca

Abstract

Kanyen’kéha (in English, Mohawk) is an Iroquoian language spoken primarily in Eastern Canada (Ontario, Québec). Classified as endangered, it has only a small number of speakers and very few younger native speakers. Consequently, teachers and courses, teaching materials and software are urgently needed. In the case of software, the polysynthetic nature of Kanyen’kéha means that the number of possible combinations grows exponentially and soon surpasses attempts to capture variant forms by hand. It is in this context that we describe an attempt to produce language teaching materials based on a generative approach. A natural language generation environment (ivi/Vinci) embedded in a web environment (VinciLingua) makes it possible to produce, by rule, variant forms of indefinite complexity. These may be used as models to explore, or as materials to which learners respond. Generated materials may take the form of written text, oral utterances, or images; responses may be typed on a keyboard, gestural (using a mouse) or, to a limited extent, oral. The software also provides complex orthographic, morphological and syntactic analysis of learner productions. We describe the trajectory of development of materials for a suite of four courses on Kanyen’kéha, the first of which will be taught in the fall of 2018.

1 Background

Kanyen’kéha (in English, Mohawk) is one of the Iroquoian[1] languages, originally spoken in the area of what is now Ontario, Québec and New York. In the period after contact, a number of Mohawk groups left or were forced to leave their homelands in New York. Initial migrations to Québec were for political and religious reasons while later migrations were forced after the Mohawk allied themselves with the British during the Revolutionary War. The refugees were provided with lands in Upper Canada. Today, there are seven Kanyen’kehà:ka communities spread across Ontario, Québec and upper New York State.

Kanyen’kéha remained the dominant language of these communities into the 20th century. However, due to a variety of well documented factors, including government assimilation policy, the use of English education in day schools and residential schools, the proliferation and domination of English media, and the desire for parents to give their children a better life, the number of first language speakers declined significantly. This precipitous decline led Hoover (1992) to write: “It is not uncommon in Kahnawake to hear people conversing with their grandchildren in Mohawk, then switching to English to speak to their own children.” In sum, Kanyen’kéha as a first language is arguably approaching the most severe level (8) on the Fishman scale (Fishman, 1991; Lewis and Simons, 2010). That is, almost all remaining L1 speakers are members of the grandparent generation, with perhaps a handful of young L1 bilinguals, the children of L2 speakers. It is difficult to determine precisely the number of remaining speakers of Kanyen’kéha, but current estimates put the number between 1000 and 1500.

Community leaders recognized the problem as it was developing and early language efforts began almost 50 years ago with night classes and the attempt to include the language in school curricula. Primary

[1] For an overview of the Iroquoian family, see Mithun (2006), and for more detail on the history of the Iroquoian languages, see Mithun (1984).

Proceedings of Workshop on Polysynthetic Languages, pages 41–52
Santa Fe, New Mexico, USA, August 20-26, 2018.

Immersion schools were started in many communities in the 1970’s and 1980’s. These were generally parent-led initiatives with community support that focused on cultural education in the language, but were not always long lasting or very effective in transmitting the language. In the 1990’s, communities shifted their focus towards creating adult speakers through full-time adult immersion programs. These programs have met with success and continue to develop as they create speakers. The central role of second language learners is also recognized. According to Stacey (2016), “the future of Kanien’kéha will depend largely upon today’s second-language speakers to become highly proficient speakers and pass the language on to their children”. Similarly, Green (2017) argues for second language instruction embedded in an immersion framework. Or again, Hoover (1992) describes “two lines of attack: a series of Mohawk classes aimed at adult non-speakers of Mohawk, and a push for community insistence on the use of the Mohawk language whenever possible.” However, these efforts are hampered by a lack of resources, including people, funding and teaching materials, as well as lack of opportunities to use the language outside the classroom.

2 Language teaching software: issues and desiderata

Along with other approaches, there have been attempts to produce language teaching software in Kanyen’kéha, including a version of Rosetta Stone (Bittinger, 2006), and a set of more advanced teaching materials including grammar exercises (Kanatawakhon-Maracle, 2002). Earlier still, collections of audio tapes, and later CD audio materials, were produced (Deering and Harries-Delisle, 2007). However, all these materials have been bedeviled by three factors.

One is the lack of academic grammars of the language, and of focussed studies of elements of the language. Many grammars are older (Bonvillain, 1973), as are many dictionaries (Bonvillain and Francis, 1972; Maracle, 2001; McDonald, 1977; Michelson, 1973). Most are not now distributed in any quantity. The same may be said of textbooks, which are relatively few and relatively difficult to obtain (Deering and Harries-Delisle, 2007; Kanatawakhon, 2013a; Kanatawakhon, 2013b).

Another factor is technological staleness: materials designed to be run on particular computer platforms become inaccessible when the platform disappears. Thus, the Rosetta Stone dataset, designed for an early version of the Rosetta software, is now not usable on more modern versions. This issue can be addressed to some degree by use of less ‘physical’ supports like websites, although web technologies also change over time (think of the disappearance of Flash).

A final, arguably more serious, factor stems from polysyntheticity itself. In the case of Kanyen’kéha, the verbal complex composed of the verb root plus a set of prefixed and suffixed forms may be extremely complex, as the following not unusual example illustrates:

(1) yaonsá:ke’

y	*a*	*onsá:*	*k*	*e*	*’*
there	conditional future	iterative	me	to be somewhere	punctual
LOC	TNS	IT	PRON	V	ASP

‘I could go back there’

Here, elements of the verbal complex are separated by spaces, although in writing they would be joined. In addition, each of the elements shown here enters into a paradigmatic relation with other possible forms within the various classes, where LOC=locative, TNS=tense, IT=iterative, PRON=pronominal, V=verb, ASP=aspect. Thus, the translocative *ye* may be replaced by the cislocative *ti*, the *a* of the conditional future by *en* for the certain future. The iterative may appear as an alternate form with a different pronominal prefix. The resulting combinatorial explosion means that the set of possible combinations cannot reasonably be assembled by hand. As a result, much current language teaching software for polysynthetic languages focusses on simpler elements like greetings, nouns, names of objects and such, despite the clear need for more complex treatments (Kell, 2014; Montour, 2012).

One element of a solution to language teaching for polysynthetic languages lies, it has been argued, in the use of a generative approach. The question is, which one? We would argue that the response must take account of the diversity of expertise required to develop and maintain electronic teaching materials. This typically includes a) computer science, b) linguistics, c) knowledge of the language, and d) pedagogical expertise. Some of these areas of knowledge may be shared across individuals, but the fact remains that any system developed must be i) as easy as possible to program, maintain and extend in its component software, including adaptation to other dialects, and potentially other languages; ii) as linguistically transparent as possible in order to 'unbind' linguistic skills (i.e. grammar writing) from programming; iii) capable of capturing as many linguistic traits of a language as possible and producing written and oral output judged acceptable by speakers; iv) as media-rich as possible, including variations in input (text, audio, images) and output (typing, clicking, oral); and v) easily usable (and adaptable) by frontline language teachers to meet their needs, so that they become more than simply consumers of tools and materials produced by others.

3 Existing generative approaches

Current approaches to generating complex verbal materials for morphologically complex languages may be roughly divided into three categories: generation of materials from pre-existing corpus data, batch generative processing, and the use of Finite State Transducers.

3.1 Generation from corpus data

An example of this approach is provided by the Arikiturri system (Itziar Aldabe et al., 2006), used for teaching Basque, a morphologically rich language. The starting point is a corpus of morphologically and syntactically analyzed sentences, marked up in XML. The software combines these with specifications of areas of focus provided by language teachers to generate test questions, including fill-in-the-blank, word formation, multiple choice, and error correction. Along with questions and expected answers, the system generates distractors by making morphological changes to the expected answers. In addition, since Basque is a free word order language, the materials generated in each question may be reordered with respect to the original corpus materials. Evaluation of materials prior to use is provided in two ways: by an 'ill-formed sentence rejector' component in the system, and by an interface which permits a language teacher to select or reject candidate questions, as well as making some basic adjustments. Aldabe et al. argue that use of the system provides efficiencies compared to manual production of examples by teachers, but also note that there exist gaps in data provided by their corpus. This is unsurprising, given that even a very large corpus is unlikely to contain all possible forms. The difficulty is aggravated in the case of Canadian indigenous languages by the lack of tagged corpora of any reasonable size.

3.2 Batch generative processing

As an alternative to corpus-based construction of exercises, Perez-Beltrachini et al. (2012) present a system called *GramEx*, based on Gardent et al. (2012), which uses Definite Clause Grammars representing the syntax and basic semantics of a language (in this case, French). These grammars are used to drive a Prolog query mechanism in order to produce a set of potential output sentences which are stored and may be queried to obtain sentences meeting desired patterns. These in turn are used to produce fill-in-the-blank and word order based exercises. Human intervention is still required, since the initial semantic specifications must be produced by hand, although a relatively small number of inputs produces a much larger set of outputs. In addition, a random sample of sentences produced by the grammar was evaluated by human experts. The system illustrates the power of a generative grammar, but the article provides few details on feedback provided to language learners, even for such primitive exercises as those presented. It is also unclear how the grammatical specifications used in the DCG for French could be adapted for Kanyen'kéha, where the potential number of inflected forms is significantly higher and where the placement of stress is context-dependent (see below). Finally, it is unclear how difficult it would be for a language teacher to add to or modify the system.

3.3 Use of Finite State Transducers

Arguably the most popular approach for dealing with morphologically rich languages is based on the use of Finite State Transducers, which may be described, in a nutshell, as rule-based devices for mapping between two sets of symbols such as, for example, an orthographic string as input and a morphological analysis as output, or vice-versa. For example in the Canadian context, FST's have been used to model nominal morphology (Snoek et al., 2014) and verb morphology (Harrigan et al., 2017) of Plains Cree, a polysynthetic language, and Harrigan et al. list a number of other similar projects. Harrigan notes as well that FST's can be applied to produce spellcheckers, paradigm generators, or components for CALL systems. Since our focus here is on CALL, we describe briefly three FST-based CALL systems, to illustrate their respective advantages and weaknesses.

Hurskainen (2009) describes a system for teaching Swahili, a language including noun classes and a complex agreement system. The system deals with morphological variation in a rule-governed way, as well as limited (concatenation based) syntax, including agreement. A learner may type some word-level input, such as a noun, and the system will analyze its morphological characteristics, indicating correctness, or an error message and a parse string. Learners may enter materials as they desire, depending on the constraints of the system, or may be taken on a 'guided tour', where they are prompted to produce increasingly more complex utterances (N, then N ADJ, and so on). Hurskainen's system has several advantages, including its encouragement of exploration by a learner, but several weaknesses as well. It appears to be limited strictly to orthographic input and output: other media like audio and images do not appear to be available. It is not clear how it could be modified without editing the grammatical representation, not a task accessible to typical instructors, and it is not clear whether it has been used in production.

Oahpa! (Antonsen, 2013; Antonsen et al., 2013) is a set of web-based language teaching materials for Northern Saami. It focusses on written activities including inflecting forms in isolation or in context, practicing simple lexical materials like numbers, and on basic question answering. Its lexicon includes basic semantic classification, phonotactic and morphophonological information, dialect information and translations to pivot languages. When the lexical database is generated, the morphological forms of words (including morphophonological variants) are also generated by the FST and saved in database tables, providing for detailed feedback on morphological errors. It is not clear, however, how that would deal with the variable placement of accented syllables in languages like Kanyen'kéha (see below). The contents of the lexicon are stored in XML, so that linguists familiar with that formalism can make changes, but it is not clear how new exercise types could be added. The results of user sessions are logged, permitting analysis of learner difficulties. It is claimed that the software is extensible to other languages.

In fact, the software behind Oahpa! has been adapted for the teaching of Plains Cree, under the title *nêhiyawêtân*. Bontogon (2016) has performed an analysis of its effectiveness using interviews with users and observation of use. While recognizing its value and potential, she flags up several issues, including interface issues (leftovers from the initial Finnish interface) and challenges to upkeep, since developers were not physically present, and it does not appear that teachers and students have the ability to change parameters or add features. To these comments, we would add the written-only character of input and output, and the still limited set of activities provided.

4 An alternative approach

In this section, we will present an alternative approach to language generation and CALL, designed to meet the desiderata mentioned in Section 2. The basic architecture is shown in Figure 1.

The three major components of the system include a natural language generator (**Vinci**) embedded in an editor (**ivi**), both written in C. The ivi editor may be used to edit both grammars and lexica. The stored language specifications are then used by Vinci to generate utterances on demand and to analyze student responses (Levison et al., 2001; Lessard and Levison, 2007). The ivi/Vinci system interacts with a user interface (**stdnt**), written in PHP, JavaScript, CSS and html5, responsible for presenting materials (written, audio, visual) to a learner within a web interface, and capturing and presenting to ivi/Vinci learner input for analysis. At the same time, a back end to the system (**instr**), written in PHP, provides grammatical and pedagogical direction to the process (what materials to produce, in what language, using what exercise format) as well as ensuring administrative tasks such as ensuring that learners are

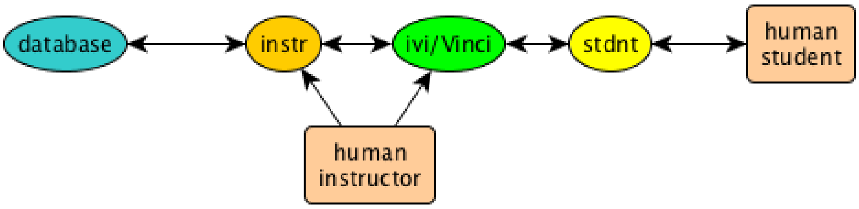


Figure 1: The architecture of the VinciLingua system

in particular classes, determining what exercises are available, and recording how individual learners have performed. The entire system is called **VinciLingua** (Lessard and Levison, 2018) and has been used successfully in production mode over the past four years by the office of Continuing and Distance Studies at Queen's University in Canada, in the context of two online French courses. Over the past year, we have been adapting the system for use in Kanyen'kéha, with a view to using it in a series of beginning University-level language and culture courses, the first of which will be taught in the fall of 2018.

The ivi/Vinci component makes use of a linguist-friendly set of metalanguages for specifying the elements of a language, including semantics, syntax, lexicon and morphology. More precisely, it is an attributed phrase structure grammar, enhanced with transformations, multipass morphological rules, and complex lexical entries, which may be preselected to drive syntactic choices. We provide here a more detailed view of how these elements work together.

4.1 Attributes and terminals

The basic building blocks of a grammar are a set of attributes and terminal symbols. The following is a simple attributes file for Kanyen'kéha:

```
PNG (p1s, p2s, p3sm, p3sf, p3si, p1d2, p1d3, p1p2, p1p3, p2d, p2p,
         p3dm, p3df, p3pm, p3pf)
Language (mohawk, english)
Meaning (action, negemotion, posemotion, poseval, negeval, state)
Medium (audio, image, text)
PrefType (ptense, pperson)
SuffType (stense)
Tense (present, past, future, conditional)
Stem (a_stem, c_stem, en_stem, i_stem, on_stem)
```

Attributes are composed of classes and values. Here, the class **PNG** represents the various persons available for pronominals in Kanyen'kéha, as in 'p1s' (first person singular), 'p2d' (second person dual), and so on. The class **Language** allows the generation of one root tree (see below) whose language is Kanyen'kéha, and another whose language is English. The **Meaning** class represents here a very basic ontology for verbs, including actions, negative emotions (for example, sadness), positive emotions, positive evaluations (think pretty, strong), and so on. Some attributes are used to control elements of morphology. Thus, **PrefType** allows the differentiation of pre-prenominal prefixes (like *en-*) for the future tense, and person pronominals (like *wak-*, 'to me'). The class **Tense** has the obvious meaning, while **Stem** allows the system to select particular classes of verbs based on their stem-type. Attributes may be partially ordered, so that, for example, both humans and animals are mobile, and compounded, as in 'action.past', to allow for dynamically richer sets. Finally, attributes may be used to control the medium of output, permitting the generation of parallel audio and textual elements, images and text, and so on.

The specification of terminals is very simple and consists in defining their labels, as in **V, PREF** and so on. Two special terminals, **BEGIN** and **END**, are used to mark the beginning and end of sequences; this is used in the dynamic calculation of accented syllables.

4.2 Lexical entries

The combination of attributes and terminals allows lexical entries to be defined. The following example shows an abbreviated form of a simple verbal entry as it would appear in the ivi editor:

```
1/HWord:  "na'khwen'on"{angry}^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
2/POS:    V^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
3/Attr:   c_stem, negemotion, PNG, Tense, Language, Medium^^^^^^^^^^^
4/Freq :  ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
5/LRule:  $v_conj^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
6/MRule:  $accentb^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
13/Prop:  cstem, p_hne, f_hake, s3, a12^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
17/Engl1: $e_conj^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
18/Engl2: "angry"^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
24/Rphn1: "na'"^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
25/Rphn2: "khwen"^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
26/Rphn3: "'on"^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
27/TBD:   ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
```

Each lexical entry represents an editable record in the lexicon, divided into fields. By definition, the first field contains the headword, followed by an English equivalent as a comment. This is followed by the part of speech (here **V**), and a set of attributes. We see that the verb is a c_stem verb which specifies a negative emotion, and that it can take all members of the class of Person, Tense and Gender, all tenses, all languages, and all media. (The use of the class in the lexicon rather than a value is the means of showing that all values are possible.)

Lexical entries may have **properties**, specific values which specify idiosyncratic elements which may be selected and used to determine choices during generation. So we see that this verb is a cstem verb, that its past tense suffix is *-hne* as opposed to other possibilities like *-hkwe* or *-hahkwe*. Information is also presented for English equivalents, including morphology rules and strings. The property **s3** indicates that this is a three syllable verb root, while the property **a12** shows that accent can occur on either the last or previous syllable (counting from end of word).

Fields 24, 25 and 26 are important. They contain, beginning in field 26 and counting backwards, the syllables which make up the verb. Division of each lexical entry into syllables is provided by an AWK script at the time of lexical entry, with subsequent human correction of the division.

4.3 Morphology rules

The Vinci system makes use of multiple pass morphology, where the output of one rule may be processed by a second rule[2]. In the case of the lexical entry above, two rules are called: **$v_conj** and **$accentb**. The first adds the basic form of the verb, and the second determines its form in context. This feature allows a morphology rule to select the appropriately accented syllable and to transform it by adding pitch and length markers, thus capturing an important feature of Kanyen'kéha: the accented syllable is determined by the number of following syllables, which depends in turn on the structure of the particular verbal complex. So in the past tense, where it is followed by *-hne*, *na'khwen'on* becomes *na'khwen'***ón***hne*, whereas in the present, it takes the form *na'k***hwén***'on*.

We will illustrate the structure of a morphology rule with a simple partial example, designed to select the appropriate form of the duals meaning 'you and I' or 'someone and I'.

```
rule yonkeni
        english: #17;
        audio: #20;
        image: #21;
        1=<13=cstem> : "yonkeni";
        1=<13=astem> : "yonky";
        ...
%
```

The rule begins by checking for the attribute 'english'. If it finds it, the contents of field 17 in the lexical entry are output (here, the rule **$e_conj** which generates the equivalent English form or sequence). This

[2] This is similar in some respects to the TWOLC model found in FST.

allows for translation. If the attribute 'audio' is found, then the appropriate sound file will be called, and if the attribute 'image' is found, the appropriate image file is called for presentation on the screen. Otherwise, output will be in Kanyen'kéha. In that case, a check is done of the properties of the following item, here the verb. (In this system, 1=next, -1=previous, -2=second previous, and so on.) If the verb is of type `cstem`, then the pronominal takes the form *yonkeni*. If the verb is of type `astem` then the pronominal takes the form *yonky*, and so on.

4.4 Syntax rules

Syntactic rules in the system take the form of context-free phrase structure rules augmented by attributes, transformations, and lexical preselections. The object of this architecture is to provide maximum delicacy of generation, while retaining readability for non-programmer linguists writing grammars, since a small change at the top of the set of rules (in the preselection) can lead to significantly different output. We provide a simple example here:

```
PRESELECT =  a : V[past, p1s]/13=cstem/$0
%
ROOT =  SELECT Pe: PNG _in_ a, Te: Tense _in_ a;
   BEGIN
     PREF[ptense, Te]  PREF[pperson, Pe, Te]
     V[c_stem, Pe, Te]/_pre_ a  SUFF[stense, Te]
   END
%
QUESTION = ROOT %
ANSWER = MAKE_ENG:ROOT  %
MAKE_ENG = TRANSFORMATION
* PREF PREF V SUFF * : 1 2[english] 3[english] 4[english] 5[english] 6;
%
```

This small syntax may be divided into four parts: a **preselection**, **definition** of the ROOT tree, **transforms** of the ROOT to define a question and an answer, and **definitions** of the transformations themselves. We will take each in turn.

The preselection finds in the lexicon all forms which match its pattern, then randomly chooses one of the forms. Here, it is looking for a past tense first person singular verb, with the property 'cstem'. The frequency operator $0 at the end specifies that once each form has been found its frequency becomes zero, thus guaranteeing that it will not be used again. This is important in questions like multiple-choice, where all options should be different. It is also possible to alter frequency in other ways, by halving it, doubling it, and so on.

Once a verb has been chosen by a preselection, its values become available to the syntax. This allows the ROOT rule to **SELECT** for use the Person and Tense, give them labels (**Pe** and **Te** respectively) and attach these to nodes on the tree. Another operator `_pre_` ensures that the verb itself in the root tree will be the one previously preselected.

The root tree itself is composed of the sequence BEGIN, PREF, PREF, V, SUFF, END, although it might in more complicated cases include branches and choices, as in any typical phrase structure grammar[3].

Note also that the terminal symbols BEGIN and END serve as markers for the attribution of accent, by providing an 'anchor' for the elements of the tree. The first PREF carries a tense marker, like *en-* to signal the certain future, the second carries the personal pronominal, like *wak-* to show the first person singular, which means, in the case of most stative verbs, something like 'to me'. The verb is obvious, while the SUFF carries a further tense marker, like *-hne*.

Specification of the ROOT is followed by creation of parallel trees for the question and expected answer. Here, the QUESTION tree is simply a copy of the root tree, while the ANSWER tree is obtained

[3] It is also important to note that it might have been possible to handle all elements of the verbal complex strictly within morphology rules, leaving syntax to handle combinations of words. We have not adopted this approach since it leads to extremely complex morphological rules, and since it complicates the option of creating parallel trees (see below). So in our system, syntax captures both the combination of separate lexical elements in an utterance, but also the structure of combined elements within the verbal complex.

by applying a transformation to the root. An indefinite number of parallel trees may be produced in order to model variant answers or to model expected errors.

Transformations (for example, here, `MAKE_ENG`) take the form of patterns and actions. Here, any sequence of `PREF PREF V SUFF`, with or without prior or following elements (indicated by the asterisks), will have the attribute 'english' attached to them, thus ensuring that the morphology rule will select the English form in the item's lexical entry.

In other words, this syntax carries the information: find a verb of such a type, make a tree by adding prefixes and suffixes, make copies of the tree, where the first copy is the question, and the second the expected answer, where the expected answer is the English equivalent of the verb.

The result of this is a series (in principle, indefinite in length) of generated questions and expected answers, as the following examples illustrate:

- Question : wakye'ónhne
  Answer : I was awake

- Question : wakerhá:rehkwe
  Answer : I was waiting

- Question : wake'nikonhren'tónhne
  Answer : I was depressed

Note that the question need not take the form of an actual question, but may be some textual, visual or audio element or set of elements presented to the learner for his or her reaction. Similarly, the answer may take the form of a written text, but images or to a limited extent sounds are also possible. Thus, one of the exercises available to learners shows pictures of animals and asks the learner to find the appropriate name. In an extension of this which we are beginning to explore, the system generates an utterance in Kanyen'kéha, for example *wake'nikonhren'tónhne* (='to me' + 'depressed' + 'past'), which forms the expected answer, and a series of stick images, which represent the meaning to be expressed: for example, a stick figure pointing at self + a sad face + an arrow pointing backwards, which form the question.

4.5 Evaluation of output

As noted earlier, the grammar and lexica for Kanyen'kéha are currently under development, for use in four courses to be taught over the next two years. It is estimated that on the order of 250 base (i.e. inflectable) forms will be used in each course, including nouns, verbal bases, prefixes and suffixes, for a total of approximately 1000 base forms. This is then not a large lexicon. However, with the inclusion of the appropriate syntactic and morphological rules, the number of potential forms rises significantly. In addition, as Snoek et al. (2014) have noted, once the basic rules have been created, addition of lexical materials becomes less complex.

For any generative system, evaluation of its (in principle indefinitely large) output must be found. In the case of VinciLingua, this is done in two ways. First, ivi/Vinci may be run in batch mode to output all forms of a structure into a spreadsheet, which may be reviewed by a human language expert. This is what is now done, and given the size of the current lexicon, it is still feasible. Second, on the language learning webpage itself, instructors may cause the system to generate sets of questions and answers and select only a subset to be shown to learners. This has been used regularly to produce quizzes in work on French.

5 Analysis of learner responses

In a software review, Dyck (2002) expresses frustration with a piece of language teaching software's requirement that a learner response match perfectly the expected pattern or be rejected. Beyond multiple choice, we would like more advanced learners to be able to type answers and have them diagnosed by the system.

One of the significant advantages of a generative approach is that it can make this possible by producing not only expected answers, but also by generating potential errors by rule, or, strictly speaking, by

malrule (Sleeman, 1982). Malrules may be thought of as rules which specify an expected error, or from another perspective, as representations in the mind of a student which produce erroneous output. The ivi/Vinci system caters for errors or variation in lexical choice, syntax, morphology, and orthography.

In the case of morphology, the system tests all variants of a morphological rule to determine whether a learner has mistakenly used the wrong variant. So in the case of the morphology rule for *yonkeni-*, all variant forms generated by the rule (*yonkeni-, yonky- yonken-*) are silently generated. If the learner’s response does not match the expected correct form, it is compared to all silently generated forms. If one matches, it is presumed to be based on a malrule, and the appropriate error message can be generated, pointing out what the student has missed.

In the case of orthography, morphology rules may capture dialectal variation, which Harrigan et al. (2017) have flagged as a significant challenge to an FST, in several ways. Thus, within Kanyen’kéha, some dialects use the letter *y* while others use the letter *i*. This may be captured by means of a morphological rule **$y**, with attributes for each dialect, to generate either the string *y* or the string *i* and to signal changes from one to the other. Alternatively, since a Vinci lexicon includes **lexical pointers** to related forms such as spelling variants, synonyms, antonyms, or other related items, the dialectal variation found in the particle for ‘something that must happen’ which takes the form *tká:konte* in Tyendinaga, *ó:nen’k tsi* in Kahnawà:ke, or *entá:onk* in Kanehsatà:ke, may be captured by an addition to the various lexical entries, as in:

```
"tká:konte"|N|action, ....|dial1:"ó:nen’k tsi"; dial2:  "entá:onk"...
```

and so on, so that if a learner enters a dialectal variant of the expected word, the system can react appropriately, either accepting the variant without comment, or flagging up its difference.

In the case of syntax, transforms of the root tree (see above) may represent not only correct but also expected erroneous structures. So, for example, missing out a pre-pronominal tense marker may be captured by a transformation like:

```
NOTENSEPREF  = TRANSFORMATION
* PREF[ptense] * : 1 3 ;
%
```

Matches to erroneous transformations can be used as the starting point for error messages to the student. Similar rules may be used to capture mis-orderings of morphemes, or addition of spurious morphemes.

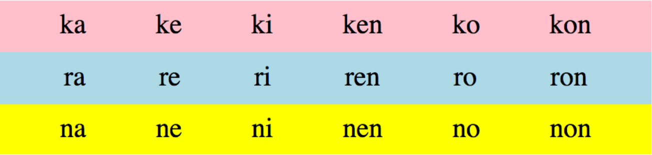
6 Multimedia and exploration

There exists some some limited evidence regarding the important role of accented syllables[4] in the acquisition of Kanyen’kéha. Thus Mithun (1989) studied five young speakers ranging in age from 1 to 5 and compared their use of complex polysynthetic forms. She found that the youngest speaker tended to retain and produce the accented syllable, responding to phonological salience; somewhat older speakers moved progressively leftward in complex forms, adding additional syllables; and the oldest speaker had reached the pronominal form and appears to be using rudimentary morphological rules.

But what of L2 learners? On the basis of classroom intuitions, we hypothesize that syllables will form an important early step and should be made as salient as possible. To capture syllables and longer oral utterances, we make use of web-based recording, followed by automatic post-processing using the **sox** software. The resulting sound files may be associated with their orthography in tables, so that a learner can click on sets to learn patterns, as Figure 2, taken from a webpage, illustrates.

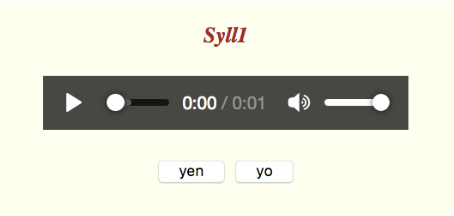
At the same time, the soundfiles for syllables may also be reused by an ivi/Vinci grammar, so that output of a set of rules is composed of one or more soundfiles played on a webpage. This permits a pedagogical trajectory, with tables being followed by simple exercises based on listening and differentiating individual syllables, as in Figure 3, more complex exercises where, given a full pronominal/verbal complex, a learner must identify meaning, as in Figure 4, or even more complex activities involving writing the equivalent of an oral utterance (in French, the ‘dictée’).

[4] Of course, accented syllables are only a small part of acquisition in polysynthetic languages. For example, Allen et al. (in press) have developed test procedures for Inuktitut which analyze multiple dimensions.



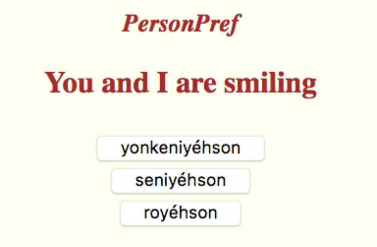
ka	ke	ki	ken	ko	kon
ra	re	ri	ren	ro	ron
na	ne	ni	nen	no	non

Figure 2: A (partial) syllable table for exploration



Syll1

Figure 3: An exercise to practice syllable differentiation



PersonPref

You and I are smiling

Figure 4: An exercise to identify verbal forms

7 Conclusions and Future Work

The crucial conclusions of the work described here are these:

- a generative approach, combined with the use of varied media, including sounds and images, represents a significant economy of effort in the production of the forms of Kanyen'kéha, and potentially for other polysynthetic languages;
- because the same grammar may underlie different activities, it is possible to define a trajectory for learners, starting with exploration of oral and written materials, to differentiation, then analysis and (at least in writing) production, while associating learner divergences from the target with tendencies to be addressed either in teaching materials or in class;
- at the same time, a generative approach, while valuable, must be included in a broader sequence of activities including cultural knowledge, texts, and activities in the language, both in and out of the classroom. This includes folktales and legends (Shakokwenionkwas, 2008; Williams, 1976), which we have not touched on here.

Acknowledgements

The work described here is a result of a partnership between Tsi Tyónnheht Onkwawén:na Language and Cultural Centre on the Tyendinaga Mohawk Territory, and Queen's University at Kingston, including the office of Continuing and Distance Studies, the School of Computing, and the Department of Languages, Literatures and Cultures, with funding provided by the Province of Ontario. We would also like to thank the anonymous reviewers for their valuable comments.

References

Shanley E. M. Allen, Catherine B. Dench, and Kerry Isakson. in press. InuLARSP: An Adaptation of the Language Assessment Remediation and Screening Procedure for Inuktitut. In M.J. Ball, D. Crystal, and P. Fletcher, editors, *Assessing grammar: Even more languages of LARSP.* Multilingual Matters, Clevedon, UK.

Lene Antonsen, Ryan Johnson, Trond Trosterud, and Heli Uibo. 2013. Generating modular grammar exercises with finite-state transducers. In *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013*, number 17 in NEALT Proceedings Series, pages 27–38. Linköping Electronic Conference Proceedings 86.

Lene Antonsen. 2013. Constraints in free-input question-answering drills. In *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013*, number 17 in NEALT Proceedings Series, pages 11–26, Linköping Electronic Conference Proceedings 86.

Marion Bittinger. 2006. Software helps revitalize use of Mohawk language. *Multilingual Magazine*, pages 59–61.

Megan Bontogon. 2016. Evaluating nêhiyawêtân: A computer assisted language learning (CALL) application for Plains Cree. Master's thesis, University of Alberta.

Nancy Bonvillain and Beatrice Francis. 1972. *A Mohawk and English Dictionary*. New York State Education Department, Albany.

Nancy Bonvillain. 1973. *A grammar of Akwesasne Mohawk*. Number 8 in Ethnology Division, Mercury Series. National Museum of Man, Ottawa.

Nora Deering and Helga Harries-Delisle. 2007. *Mohawk: a teaching grammar*. Kanien'kehá:ka Onkwawén:na Raotitióhwka Language and Cultural Center, Kahnawàke, 2nd edition.

Carrie Dyck. 2002. Review of Tsi Karhakta: At The Edge of the Woods. *Language Learning & Technology*, 6(2):27–33.

Joshua A. Fishman. 1991. *Reversing language shift: theoretical and empirical foundations of assistance to threatened languages*. Multilingual Matters, Clevedon.

Claire Gardent and German Kruszewski. 2012. Generation for Grammar Engineering. In *INLG 2012, The seventh International Natural Language Generation Conference*, pages 31–40.

Jeremy Green. 2017. Pathways to creating Onkwehonwehnéha speakers at Six Nations of The Grand River Territory. Technical report, Six Nations Polytechnic.

Atticus G. Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27:565–598.

Michael Hoover. 1992. The revival of the Mohawk language in Kahnawake. *Canadian Journal of Native Studies*, 12(2):269–287.

Arvi Hurskainen. 2009. Intelligent Computer-Assisted Language Learning: Implementation to Swahili. Technical report, Institute for Asian and African Studies; University of Helsinki, http://www.njas.helsinki.fi/salama.

Maddalen Lopez de Lacalle Itziar Aldabe, Montse Mar-itxalar, Edurne Martinez, and Larraitz Uria. 2006. Arikiturri: an automatic question generator based on corpora and NLP techniques. In *Proceedings of the 8th international conference on Intelligent Tutoring Systems, ITS'06*, pages 584–594, Berlin, Heidelberg. Springer-Verlag.

David Kanatawakhon-Maracle, 2002. *Tsi Karhakta: At The Edge of the Woods (Mohawk courseware).*

David Kanatawakhon. 2013a. *To I'i Tewaweyentehta'n ne Kanyen'keha. Let's learn Mohawk: an introductory grammar text for learning the Mohawk language*. Centre for Research and Teaching of Canadian Native Languages, University of Western Ontario, London, ON.

David Kanatawakhon. 2013b. *To I'i Tewaweyentehta'n ne Kanyen'keha. Let's learn Mohawk: a text of grammar supplements concerning nominals*. Centre for Research and Teaching of Canadian Native Languages, University of Western Ontario, London, ON.

Sarah Kell. 2014. Polysynthetic Language Structures and their Role in Pedagogy and Curriculum for BC Indigenous Languages: Final Report. Technical report, British Columbia Ministry of Education.

Greg Lessard and Michael Levison. 2007. Lexical creativity in L2 French and Natural Language Generation. In Dalilah Ayoun, editor, *French Applied Linguistics*, pages 299–333. John Benjamins.

Greg Lessard and Michael Levison. 2018. Vincilingua website. `https://vincilingua.ca`.

Michael Levison, Greg Lessard, Anna Marie Danielson, and Delphine Merven. 2001. From Symptoms to Diagnosis. In Keith Cameron, editor, *CALL - The Challenge of Change*, pages 53–59.

M. Paul Lewis and Gary F. Simons. 2010. Assessing Endangerment: Expanding Fishman's Grids. *Revue Roumaine de Linguistique*, LV(2):103–120.

David Kanatawakhon Maracle. 2001. *Mohawk Language Thematic Dictionary*. Kanyen'keha Books, London, ON.

Mary McDonald. 1977. *Iontenwennaweienstahkhwa': Mohawk Spelling Dictionary*, volume Bulletin 429. New York State Museum, Albany.

Gunther Michelson. 1973. *A thousand words of Mohawk*. Number 5 in Ethnology Division, Mercury Series. National Museum of Man, Ottawa.

Marianne Mithun. 1984. The Proto-Iroquoians: Cultural reconstruction from lexical materials. In Jack Campisi, Michael K. Foster, and Marianne Mithun, editors, *Extending the Rafters*, pages 259–282. SUNY Press, Albany.

Marianne Mithun. 1989. The acquisition of polysynthesis. *Journal of Child Language*, 16:285–312.

Marianne Mithun. 2006. The Iroquoian languages. In *Encyclopedia of Language and Linguistics*, volume 6, pages 31–34. Elsevier, Oxford, 2nd edition.

Barry M. Montour. 2012. The Kanien'kéha Proficiency Assessment. In *First Nations Lifelong Learning Assessment Report*, pages 22–27.

Laura Perez-Beltrachini, Claire Gardent, and German Kruszewski. 2012. Generating Grammar Exercises. In *The 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 147–156. Association for Computational Linguistics.

Tom Porter Shakokwenionkwas. 2008. *And Grandma Said... Iroquois Teachings*. Private publication, Kanatsiohareke Mohawk Community, Fonda, New York.

Derek Sleeman. 1982. An attempt to understand students' understanding of basic algebra. *Cognitive Science*, 8(4):387–412.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the Noun Morphology of Plains Cree. In *Proceedings of the 2014 workshop on the use of computational methods in the study of endangered languages*, pages 34–42. Association for Computational Linguistics.

Kahtehrón:ni Iris Stacey. 2016. Ientsitewate'nikonhraié:ra'te tsi nonkwá:ti ne á:se tahatikonhsontóntie: We will turn our minds there once again, to the faces yet to come. Master's thesis, University of Victoria.

Marianne Williams, editor. 1976. *Kanien'kéha' Okara'shón:'a (Mohawk Stories)*, volume 427, Albany. New York State Museum.