Prototype d'analyse automatique pour le kanyen'kéha : langue polysynthétique iroquoienne

Nathan Brinklow[1], François Daoust[2], Monique Dufresne[1], Greg Lessard[1], Rose-Marie Déchaine[3]

[1]Queen's University – nathan.brinklow@queens.ca, dufresne@queensu.ca, lessard@cs.queensu.ca
[2]Université du Québec à Montréal – daoust.françois@gmail.com
[3]University of British Columbia – dechaine@mail.ubc.ca

Résumé

Le kanyen'kéha (mohawk), langue iroquoienne parlée aux Canada et aux États-Unis, se caractérise par une morphologie complexe qui génère un ensemble important de formes verbales et nominales. Par exemple, l'expansion minimale d'une forme verbale est bi-morphémique, comprenant une racine et un préfixe pronominal (p.ex. ***wak****yó'te* '**je** travaille') et fait partie d'un paradigme de 90 cellules : 15 pronoms x 6 radicaux. Ainsi un des grands défis est la reconnaissance d'une racine, celle-ci étant entourée d'un nombre variable et parfois élevé d'affixes : pour le schéma verbal, l'expansion maximale donne au moins 14 positions affixales (5 préfixes, 9 suffixes). Dans ce contexte et avec le but de développer un répertoire de méthodes et d'outils d'analyse plus performants, et en s'appuyant sur les données d'un corpus du kanyen'kéha, nous présentons une passerelle entre deux logiciels – l'un axé sur une analyse inductive (*Système d'analyse de textes par ordinateur, SATO*) et l'autre axé sur la génération déductive (*ivi*/*Vinci*).

Mots clés : Kanyen'kéha, mohawk, iroquoienne, polysynthèse, analyse de texte, génération automatique, segmentation morphologique, *SATO*, *ivi*/*Vinci*.

Abstract

Kanyen'kéha (Mohawk) — an Iroquoian language spoken in Canada and the U.S.A. — is characterized by complex morphology that derives a large set of verbal and nominal forms. For example, the minimal expansion of a verb form is bi-morphemic, and includes a stem and a pronominal prefix, e.g. ***wak****yó'te* '**I** work'. Even restricting attention to such bi-morphemic forms, this generates a paradigm space with 90 cells: 15 pronouns x 6 stem types. In addition, for a given stem, recognition of the root poses a challenge, as the latter is surrounded by a variable and sometimes large number of affixes, with the maximal expansion of the verb template containing at least 14 affix positions (5 prefixes, 9 suffixes). In this context, we develop a powerful repertoire of analytic methods and tools for the treatment of Kanyen'kéha corpus data. We do so by bridging two computational environments: one based on inductive analysis (*Système d'analyse de textes par ordinateur, SATO*) and the other based on deductive generation (*ivi*/*Vinci*).

Keywords: Kanyen'kéha, Mohawk, Iroquoian, polysynthesis, text analysis, automatic generation, morphological segmentation, *SATO*, *ivi*/*Vinci*.

1. Le défi des langues vulnérables

Toutes les langues autochtones des Amériques sont vulnérables et dans bien des cas, les enfants ne les apprennent plus comme langue première. Le kanyen'kéha, langue iroquoienne parlée au Canada et aux États-Unis, fait partie des langues les plus à risque et les plus menacées en Amérique du Nord. On estime qu'il existe environ 2 000 locuteurs qui le parlent couramment, dont environ 900 locuteurs adultes L1, 1000 locuteurs adultes L2 et 30 enfants L1 (Green 2018). À l'heure actuelle, il n'existe que peu de ressources pour les apprenants avancés, ressources qui

leur permettraient d'être novateurs, soit en créant de nouvelles formes langagières, soit en analysant des formes jamais rencontrées. Pour pallier cette lacune et avec le but de développer des activités pédagogiques interactives, nous sommes à tester des prototypes de génération et d'analyse linguistique en intégrant deux outils informatiques. Comme première étape, nous présentons ici des algorithmes pour le balisage et l'analyse automatique en intégrant deux ensembles d'outils de calcul, à savoir le *Système d'Analyse de Texte par Ordinateur* (*Sato*) développé par F. Daoust à l'Université du Québec à Montréal et le logiciel *ivi/Vinci* développé par M. Levison et G. Lessard à l'Université Queen's.

Notre but est de concevoir et de livrer un premier corpus textuel en kanyen'kéha convivial et annoté, accessible au public sur le web. Dans cette perspective, notre recherche collaborative s'appuie sur la linguistique computationnelle, la linguistique de corpus, la pédagogie des langues et la grammaire du kanyen'kéha. Nous poursuivons quatre objectifs interdépendants : la modularisation d'outils de calcul afin de permettre des représentations multiples dans l'environnement *Sato* ; le développement et l'évaluation d'algorithmes pour l'étiquetage de parties de discours du kanyen'kéha ; l'utilisation des outils développés pour mener des recherches inductives (analyse) et déductives (génération) sur corpus ; et l'accroissement de la conscience linguistique de l'apprenant sur (a) les patrons paradigmatiques via la génération de nouvelles formes langagières et (b) les styles discursifs via l'analyse de formes langagières.

2. Les défis posés par la complexité morphologique du kanyen'kéha

2.1. La complexité combinatoire en kanyen'kéha

À l'instar de toutes les langues iroquoiennes, le kanyen'kéha présente une morphologie flexionnelle et dérivationnelle très riche (Mithun 2006). Ainsi un verbe, entouré de ses affixes, constitue à lui seul l'équivalent informationnel d'une phrase entière dans une langue comme le français ou l'anglais. L'exemple (1) en fournit une illustration où il est possible d'identifier les six morphèmes suivants : (i) le radical verbal ***-stikawh-*** 'errer' ; (ii) le préfixe duplicatif (DUP) ***te-*** ; (iii) le préfixe pronominal 3e personne masculin singulier ***ho-***, (iv) le préfixe réfléchi (REFL) ***t-*** ; (v) le suffixe aspectuel perfectif (PFV) ***-en*** ; (vi) le suffixe aspectuel progressif (prog) ***-hákye*** et (vii) le suffixe ponctuel ***-'***[1].

(1) *tehotstikawhenhákye'*
te-ho-t-stikawh-en-hákye-'
DUP-3SG.M-REFL-errer-PFV-PROG-PFV GLOSE1 (Bonvillain 1970)
DUP-3SG.M-REFL-errer-PFV-PROG-PONCTUEL GLOSE2 (GRAM)
'il errait' (adapté de Bonvillain 1970 : 86, (1))

Sur le plan linéaire, la complexité de ce système se manifeste dans le schéma en (2), qui comprend six positions, dont : (***a***) la marge gauche pré-préfixale qui présente une interpolation de mode et d'aspect (Foster 1985; 1986) ; (***b***) les préfixes pronominaux; (**c**) les noms incorporés (incluant les réfléchis) ; (**d**) la base verbale ; (**e**) les suffixes de la valence verbale (INCHOATIF, CAUSATIF, APPLICATIF) ; (**f**) les suffixes de l'aspect grammatical (PONCTUEL, HABITUEL, STATIF).

[1] Conventions : 1 = 1re personne, 3 = 3e personne, 1:3 = 1re personne sur 3e personne (transitif), ASP = aspect, ATTRIBUT.POS = attribut positif, BEN = bénéfactif, CSV = causatif, DEF = défini, DIST = distributif, DUP = duplicatif, FM = féminin, IRR = irréalis, ITER = itératif, LOC = translocatif, M = masculin, NMLZ = nominalisation, N = nom, NT = neutre, PC = particule, PFV = perfectif, PL = pluriel, PNC = ponctuel, PPREF = préfixe pronominal, PREF = préfixe verbal (à l'exclusion des préfixes pronominaux), PROG = progressif, PURP = intentionnel, REFL = réfléchi, SUFF = suffixe, V = verbe, SG = singulier, ZC = zoic.

(2)	*a-*	*b-*	*c-*	*d*	*-e*	*-f*
	pré-	préfixes	noms	base	suffixes	suffixes
	préfixes	pronominaux	incorporés	verbale	de valence	aspectuels

Les racines nominales impliquent l'existence de plusieurs allomorphes dont la sélection est partiellement régie, soit l'utilisation du nom comme nom indépendant ou nom incorporé. Le nom indépendant se présente selon sa classe sous une forme tronquée où les suffixes dérivationnels sont souvent absents (symptôme d'une lexicalisation) (cf. Barrie & Jung 2020) ; voir la forme tronquée en (3a) et une forme plus longue en (3b). Lorsque le nom est incorporé, les formes longues sont sélectionnées : avec nominalisation (-*'sere-'tsher-*) ou sans nominalisation (-*'sere-'t-)* ; le dernier est illustré en (3c-d).

(3) a. *kà:sere*
ka-'sere
3SG.NT.ZC-trainer
'voiture'

b. *ka'**sere'tshér**a*
ka-'sere-'t-sher-a
3SG.NT.ZC-[trainer]-CSV-NMLZ- FINALE
'voiture, lit., 'il le fait trainer'

c. *ka'**sere't**í:yo*
ka-'sere-'t-sher-iyo
3SG.NT.ZC-[trainer]-CSV-NMLZ- ATTRIBUT.POS
'C'est une bonne voiture.'
lit. 'c'est une bonne qui-se-fait-trainer'

d. *ke'**sere-t**óhares*
*ke-'**sere't-**óhare-s*
1SG-[voiture-laver]-HAB
'Je lave la voiture.'
lit. 'je lave ce-qui-se-fait-trainer'

2.2. L'allomorphie en kanyen'kéha

Les possibilités combinatoires du kanyen'kéha sont davantage compliquées par l'allomorphie. Comme on le voit en (4) le même préfixe pronominal de 3e personne masculin prend deux formes distinctes selon le contexte phonologique : *ra-* apparait en position initiale, *ha-* se trouve en position interne.

(4) a. ***rati**khón:nis*
ra-ti-khón:ni-s
3MS-PL-[cuire]-HABITUEL
'ils cuisent' (actuellement, tout le temps, comme métier)

b. *en**hati**khón:ni'*
en-**ha-ti**-khón:ni-'
IRR-3MS-PL-[cuire]-PONCTUEL
'ils vont cuire'

Une autre complication résulte du fait que le choix d'un préfixe peut déterminer la sélection d'un suffixe. Ainsi, en (4a), au présent, le suffixe habituel *-s* est possible. Par contre en (4b) l'apparition du préfixe irréalis *en-* amène la sélection du suffixe ponctuel *-**'***.

Finalement la fusion de deux morphèmes en une seule suite opaque phonologique présente un défi pour le découpage morphologique. Par exemple, en (5b) le préfixe *a-* et le préfixe pronominal *wak-* sont fusionnés pour former *aon-*. (Cf. Beatty 1985 : 85.)

(5) a. *en**wak**yó'teke*
en-**wak**-yó'te-ke
IRR-1-[travailler]-ASP
'je vais travailler'

b. *aon**k**yó'teke*
aon-**k**-yó'te-ke
IRR -1-[travailler]-ASP
'je pourrais/devrais travailler'

3. Enjeux méthodologiques

La richesse morphologique du kanyen'kéha pose des problèmes pour l'enseignement et l'apprentissage de la langue, ainsi que pour l'analyse computationnelle.

3.1. Enjeux pour la pédagogie du kanyen'kéha

Les apprenants du kanyen'kéha sont appelés à maitriser un système morphologique riche, un défi de taille. Le nombre important de combinaisons potentielles a comme conséquence que les manuels d'enseignement et la plupart des didacticiels se contentent de présenter quelques exemples illustratifs et visent surtout les expressions figées et les noms, où les possibilités combinatoires sont moindres. Il y a donc un urgent besoin pour la constitution d'environnements computationnels qui permettent la génération de paradigmes complets. De surcroit, les apprenants plus avancés font face à des défis importants puisqu'une bonne partie de l'acquisition de langue passe par la lecture et qu'il y a peu de textes en kanyen'kéha disponibles en ligne dans un format convivial.

3.2. Enjeux pour l'analyse textuelle du kanyen'kéha

L'orthographe du kanyen'kéha n'étant pas fixée, cela soulève un enjeu de taille pour l'analyse de textes. Malgré le fait que le système d'écriture du kanyen'kéha ait été standardisé en 1993 par *The Mohawk Language Steering Committee*, aucune norme orthographique n'a été imposée, le comité sentant le besoin de respecter les différences dialectales ou idiolectales. Le système d'écriture standardisé est largement répandu et les caractères suivants ont été adoptés, soient onze caractères latins : ***a***, ***e***, ***h***, ***i***, ***k***, ***n***, ***o***, ***r***, ***s***, ***t***, ***w*** ; ***y*** ou ***i*** selon le scripteur pour représenter la semi-voyelle /j/ ; l'apostrophe (**'**) pour transcrire l'occlusive glottale ; les deux points (**:**) pour indiquer la longueur d'une voyelle ; et l'accent aigu (´) et l'accent grave (`) pour indiquer une courbe intonatoire ascendante ou descendante associée à la mélodie vocalique.

Notons que malgré la standardisation, les textes rédigés depuis 1993 comportent encore beaucoup de variations, ce qui présente un défi pour l'analyse informatique. Ainsi, on relève des frontières de morphèmes rendus opaques par certaines conventions orthographiques, soit l'utilisation de l'accent grave (`) et les deux points (**:**). En (6), la segmentation du premier morphème ***ka-*** implique la disparition de l'accent grave et la substitution des deux points par l'apostrophe (**'**).

(6) *kà:sere*
ka-'sere
3SG.NT.ZC-trainer
'voiture'

L'opacité de la forme en (6) est attribuable à l'interaction entre la morphologie et la phonologie D'un point de vue computationnel, la séquence V:C peut facilement être décomposée par les formules en (7) et (8) pour la substitution du coup de glotte (**'**) ou l'aspiration (**h**), respectivement[2].

(7) $\text{ORTHO}\,(\text{V} + \grave{} + \textbf{:} + \text{C}_{\text{OCCLUSIVE}}) = \text{SEG}(\text{V} + \textbf{'} + \textbf{-} + \text{C}_{\text{OCCLUSIVE}})$

	à:C			*è:C*		*ì :k*		*ò:k*		*èn:k*		*òn:k*	
ORTHO	*à:k*	*à:t*	*à:s*	*è:k*	...	*ì:k*	...	*ò:k*	...	*èn:k*	...	*òn:k*	...
SEG	a'-k	a'-t	a'-s	e'-k	…	i'-k	…	o'-k	…	en'k	…	on'k	…

[2] Seules les voyelles qui portent un accent grave dans l'orthographe sont sujettes à une ré-analyse pour la segmentation morphologique. L'accent aigu (´) est régi phonologiquement par la présence d'une consonne ou voyelle dans une syllabe accentuée.

(8) ORTHO (V + ` + **:** + $C_{SONANTE}$) = SEG(V + h + **-** + $C_{SONANTE}$)

	à:C				*è:C*		*ì :k*		*ò:k*		*èn:k*		*òn:k*	
ORTHO	*à:n*	*à:r*	*à:w*	*à :y*	*è:n*	...	*ì:n*	...	*ò:n*	...	*èn:n*	...	*òn:k*	...
SEG	ah-n	ah-r	ah-w	ah-y	eh-n	…	ih-k	…	oh-k	…	eh-n	…	oh-n	…

D'autre part, il existe des documents, de valeur inestimable pour une analyse de corpus mais qui soulèvent de nombreux problèmes, car ils sont rédigés dans différentes orthographes influencées par le latin, le français ou l'anglais, voire dans un système syllabique. Notons en outre les textes les plus anciens rédigés en kanyen'kéha, *Radices verborum iroquaeorum,* qui datent de la fin du 17e siècle, ainsi qu'un nombre important de textes chrétiens traduits en kanyen'kéha au cours des 18e et 19e siècles. À cela s'ajoute un autre enjeu, soit la piètre qualité d'impression qui rend à toutes fins impossibles la reconnaissance optique de caractères. L'existence de plusieurs systèmes d'encodage pose un autre problème pour le rendu de diacritiques et pour la reconnaissance optique de caractères[3].

3.3. Enjeux pour l'analyse computationnelle du kanyen'kéha

La plupart des outils computationnels disponibles peinent à traiter les phénomènes énumérés ci-dessus, puisqu'ils ont été développés pour le traitement de langues morphologiquement pauvres. De plus, le petit nombre de textes en kanyen'kéha écrits accessibles en ligne constitue un écueil de taille pour l'analyse de corpus. La combinaison de deux catégories d'outils — à savoir un système de génération et un analyseur de textes — offre une solution à cette impasse. Le système de génération permet de créer des paradigmes exhaustifs, tâche difficile et ardue pour un être humain, et l'analyseur de textes rend possible la représentation d'un même corpus sur plusieurs niveaux, allant d'une analyse linguistique très fine à une représentation qui convient à la lecture des textes en ligne. Grâce à l'apport des ressources visuelles et sonores qu'offre un environnement *html5*, des informations pédagogiques plus riches sur chacun des éléments sont dès lors disponibles aux utilisateurs.

Dans ce qui suit, nous présentons un premier aspect de cet outil combiné, via une passerelle qui permet de relier les deux approches. Cette présentation prend la forme suivante : dans la section (4), il sera question de la constitution du corpus, dans la section (5), du système de génération, dans la section (6), de l'analyseur de texte, dans la section (7) des résultats d'une tentative de relier les deux, et dans la section (8) nous proposerons une première série de conclusions provisoires.

4. Constitution d'un corpus kanyen'kéha

À l'heure actuelle, des corpus annotés linguistiquement dont l'existence est importante tant pour la recherche que pour l'enseignement, font crucialement défaut. Pour combler ce besoin, on a choisi une approche prototypale permettant de tester des outils informatiques et des chaines de traitement dont le déploiement futur devrait permettre de constituer des corpus annotés avec les moyens de les exploiter. *The Bear and the Fox*, récit recueilli à Akwesasne au Québec en juin 1970 servira de corpus témoin. Ce texte et son analyse sont intéressants à maints égards, car ils permettent de :

- cibler des **phénomènes linguistiques** qui relèvent du discours, à savoir l'alternance entre

[3] Comme le notent Little *et al*. (2018) : « An unexpected problem with integrating spell-checkers into mainstream office software is tokenization, since some Indigenous languages use commas, colons, and apostrophes to indicate phonetic differences, whereas many text processing systems assume internally that these are token boundaries. This points to a need for more flexible tokenization within mainstream office software to accommodate these languages. »

les noms incorporés et non incorporés ; l'utilisation des déterminants et des particules discursives ; la manipulation du point de vue via l'alternance du discours rapporté et la narration ; et l'organisation de structure informationnelle, tel que le topic et le focus ;
- développer du **matériel pédagogique** qui peut être utilisé par et pour les apprenants avancés et qui leur permet de d'acquérir des compétences sur les structures discursives du kanyen'kéha ;
- tester le **potentiel de nos logiciels** sur un texte complexe ;
- tirer profit de la **grammaire** du kanyen'kéha formalisée par ***ivi/Vinci***, qui sera mise à contribution pour inspirer le découpage en morphèmes et leur catégorisation ; et
- tirer profit de **l'architecture de *Sato*** qui permet de gérer plusieurs couches concurrentes d'annotation (p. ex. transcription, identification des morphèmes, des grilles de catégories linguistiques et valorisation de l'annotation) qui peuvent être ajustées selon le public cible.

L'informatisation de ce corpus témoin servira donc de banc d'essai pour la mise au point de règles de segmentation et de catégorisation automatiques dont les résultats pourront être confrontés à l'analyse manuelle de Bonvillain (1970) [4].

5. *ivi/Vinci*

Le balisage des éléments est sans aucun doute l'aspect le plus long et fastidieux de l'analyse d'un corpus. Ce travail peut être accompli de façon entièrement manuelle : un être humain ajoute les balises nécessaires à chaque élément du corpus. Sauf dans les cas les plus simples, une telle approche n'est pas pratique. Une deuxième approche, plus puissante, capte dans un 'dictionnaire' la liste des formes dans un texte, de telle sorte que chaque forme n'est balisée qu'une seule fois. Comme nous le verrons dans la section suivante, c'est un des avantages offerts par le logiciel *Sato*. Mais la couverture d'une telle approche dépend du corpus traité. Si une forme est absente du corpus et du dictionnaire, elle ne sera pas prise en compte. Dans le cas du kanyen'kéha, avec sa vaste panoplie d'affixes, et en l'absence d'un nombre suffisant de corpus analysés, cette difficulté est particulièrement sérieuse.

Pour pallier cette lacune, il est possible d'utiliser une approche générative, où une grammaire produit de façon déductive et systématique l'ensemble des formes de chaque paradigme d'affixes ou de racines. Dans ce qui suit, nous illustrerons cette approche au moyen de quelques exemples traités par le logiciel *ivi/Vinci*. Sur le plan théorique, *ivi*/*Vinci* prend la forme d'une grammaire non-contextuelle de Type 2, selon la hiérarchie de Chomsky (1959), augmentée d'attributs (Knuth 1968), de conditions (Boyland 1996), et de transformations (Chomsky 1965), le tout mis en œuvre sous forme d'un ensemble de métalangages et d'opérations dans le système de génération *Vinci*, lui-même imbriqué dans un éditeur (*ivi*). Sur le plan pratique, une grammaire minimale est composée des modules suivants: (i) les éléments terminaux (N, V, etc.); (ii) les attributs sémantiques et grammaticaux (singulier, indéfini, etc.) répartis en classes, chacune composée d'un ensemble de valeurs complémentaires; (iii) les règles syntaxiques comprenant des règles de base, des conditions de développement d'arbres, et des transformations; (iv) les règles morphologiques (potentiellement récursives et sensibles au contexte); (v) les entrées lexicales regroupant chacune sa classe terminale, ses attributs, ses traits sémantiques et formels, ses règles morphologiques, et ses formes de base.

[4] *The Bear and the Fox* a fait l'objet d'une analyse linguistique par Bonvillain et Francis. Le texte se présente sous deux versions. La première comprend une transcription phonémique, les gloses pour chacun des morphèmes et finalement une traduction libre. La deuxième version présente une segmentation morphémique; une identification où chaque morphème est glosé et où la fonction des morphèmes grammaticaux est identifiée (les particules discursives échappent souvent à cette règle); et une traduction mot à mot.

La génération des éléments d'un paradigme prend la forme suivante : un patron syntaxique est créé et appelle les éléments terminaux, les unités lexicales, les règles morphologiques, etc. pour former une suite dans la langue à générer. Ainsi, le patron syntaxique en (9) produit le préfixe pronominal statif et objectif de la première personne *wake-*. L'ajout d'autres règles produit les variantes morphologiques de *wake-* , dont {*wak-*, *k-*, …}.

(9) ROOT = PPREF[p1s, stative, obj]

Quant à la règle (10), elle utilise la variable **Pe** pour représenter les différentes valeurs du trait 'personne' en kanyen'kéha, afin de produire tous les préfixes pronominaux statifs et objectifs, comme ceux de 2e personne *sa*, et de 3e personne masculin *ro*, etc., ainsi que leurs variantes contextuelles. Des règles analogues permettront la génération d'autres paradigmes d'affixes ou de racines. Pour rendre les résultats de la génération utilisables par *Sato*, des règles de transformation sont appliquées aux règles syntaxiques de base pour produire les balises associées à chaque forme. L'exemple (11) montre une balise typique, dans le format d'exportation vers un dictionnaire *Sato*. Dans les cas relativement fréquents où une même forme peut avoir plus d'une valeur, comme ***k***, qui peut représenter, entre autres, le préfixe pronominal actif dans certains contextes, ou statif dans d'autres, les différentes valeurs sont entrées dans une même règle et séparées par une virgule.

(10) ROOT = CHOOSE Pe:PNG; PPREF[Pe, stative, obj]

(11) wake*gramr=("PPREF:p1s:stative:obj")

L'utilisation des capacités d'*ivi/Vinci* permet ainsi de pré-remplir le dictionnaire *Sato*, réduisant ainsi une partie de la tâche de balisage. Et de façon complémentaire, le produit du balisage minutieux d'un corpus au moyen de *Sato* est versé dans la grammaire *ivi/Vinci* pour affiner et étendre les règles de génération.

6. Sato

Le logiciel *Sato* se présente comme une plateforme informatique destinée à soutenir des démarches d'analyse textuelle en respectant le cadre théorique de l'analyste. Le logiciel permet d'enrichir un corpus de recherche par l'ajout dynamique de systèmes d'annotations, appelés *propriétés*, qui se superposent au texte d'origine dans sa dimension lexicale et contextuelle. L'affectation de valeurs à ces propriétés peut se faire par manipulation directe à l'écran ou par des manipulations résultant de calculs faisant appel à du filtrage, à des patrons positionnels en contexte ou à des dictionnaires.

Une syntaxe de filtrage, fournissant une forme simplifiée d'expressions rationnelles sensibles aux annotations et aux chaines de caractères, permet en effet de sélectionner des ensembles de formes lexicales et/ou d'occurrences de ces formes. Ces sélections servent d'entrées aux divers algorithmes d'analyse et d'annotation. Elles servent notamment à définir des sous-textes taillés sur mesure avec la fréquence spécifique du vocabulaire utilisé. Des algorithmes textométriques simples permettent de révéler les différences entre ces sous-textes. Des décomptes en contexte sont aussi possibles pour dépister des régularités syntaxiques et discursives.

Divers formats d'exportation du corpus annoté et des résultats compilés permettent de faire appel à des logiciels externes, notamment pour du traitement linguistique automatique, pour du traitement statistique et de représentation visuelle.

Le logiciel fournit un mécanisme systématique de journalisation des manipulations sur le corpus et permet de rédiger des scénarios de commandes facilitant la documentation des étapes de l'analyse et leur reproduction. Le logiciel reposant sur une interface Web, l'analyse peut se

faire en ligne et on peut facilement fournir par sa suite un accès aux corpus annotés, en particulier à des fins de formation et de consultation par un vaste public. On peut d'ailleurs compléter l'interface Web de recherche par des pages HTML supplémentaires destinées à faciliter cette consultation élargie. Un simple clic sur un mot ou une forme lexicale permet de révéler l'information cumulée, de naviguer entre les contextes et les formes lexicales et d'ajouter, si nécessaire, ses propres annotations. L'affichage du texte et du lexique est totalement configurable pour créer des éditions sur mesure.

Il existe déjà des ressources informatiques pour l'annotation et la génération morphologique, la manipulation des bases de données lexicales, et l'analyse de corpus. Sans être exhaustif, citons UniMorph (Kirov et al., 2018) et l'environnement Fieldworks (Baines, D., 2018). UniMorph permet l'analyse et la génération de formes fléchies à partir d'un dictionnaire et d'un ensemble de règles flexionnelles. Fieldworks se définit comme un 'écosystème' de logiciels qui permet la collecte et à la manipulation des données lexicales et culturelles, l'analyse lexicographique et morphologique (au moyen des parseurs Xample et Hermit Crab), la manipulation des corpus, et la publication de dictionnaires sur le web ou sur papier. Malheureusement, comme UniMorph, il ne comprend pas de données en kanyen'kéha.

Sato possède plusieurs éléments en commun avec Fieldworks: il permet la manipulation de lexiques et de corpus et leur présentation en ligne, au moyen de l'annotation manuelle d'entrées lexicographiques et de corpus ainsi que la projection d'un dictionnaire et des règles morphologiques sur un corpus. Par contre, contrairement à Fieldworks, qui exige un environnement Windows, SATO fonctionne sur serveur à partir d'une interface web. En outre, il permet plusieurs niveaux de présentation selon l'usager (apprenants, spécialistes, etc.) et, comme nous le verrons plus bas, la création et le stockage de calculs contextuels pour baliser un élément selon sa position par rapport à d'autres éléments dans un corpus, non seulement dans le même mot, mais aussi dans la même phrase ou paragraphe.

6.1. Prototype de chaine de traitement

L'approche inductive en analyse textuelle repose sur l'explicitation des composants linguistiques tels qu'ils se présentent dans le texte plein soumis au lecteur. Pour des langues comme le kanyen'kéha, les textes sont généralement le résultat de transcriptions écrites de récits issus de la tradition orale. Pour faciliter la comparaison entre les textes du corpus, des décisions de normalisation doivent être prises dès cette première étape de l'analyse. La trace de ces décisions doit être conservée pour permettre de revenir en arrière si nécessaire.

Voici un exemple présentant la première phrase de notre prototype de corpus, tel que soumis à *Sato* dans sa version anglaise. Il s'agit d'une mise en forme de la version imprimée du corpus qui présente les annotations sur diverses lignes.

Ce premier état du corpus utilise deux propriétés lexicales : *txt* et *eng* qui contiennent respectivement la graphie du kanyen'kéha utilisée en 1970 et la traduction anglaise du mot, qui est différente de la graphie moderne, qui utilise l'alphabet nommé ici *ka*. *{…} désigne un commentaire et *Page=/77 donne le numéro de page dans l'édition imprimée du document. La section qui se termine par *Title* définit les règles de codification. Les caractères accentués, les deux points '**:**' et '' (ou '**?**' dans l'orthographe utilisée par Bonvillain) sont des caractères qui font partie de l'alphabet kanyen'kéha (et non des marques de ponctuation). Une fois soumis à *Sato*, on configure le logiciel avec un affichage convivial masquant les propriétés *txt* et *eng.* Et on y ajoute la propriété lexicale *Seg* à laquelle on affecte, pour chacune des formes, une représentation segmentée en morphèmes.

Figure 1 : Mise en forme (*Sato*)

Alphabet ka ,0 .0 ,1 .1 ,2 .2 ,3 .3 ,4 .4 ,5 .5 ,6 .6 ,7 .7 ,8 .8 ,9 .9 *séparator , . ; ! ? ... … < > ( ) [ ] { } « » ” “ % $ # " @ & + = / \ | * — ¦
Property txt free for lexicon
Property eng free for lexicon
Title The Bear and the Fox in Akwasasne Mohawk (Bonvillain and Francis)
*{The following tale, 'The Bear and the Fox,' was collected from Ms. Beatrice Francis of Akwesasne, Quebec, during several sessions in June 1970. It is Francis' Mohawk translation of the folktale recorded by Floyd Lounsbury in his Oneida Verb Morphologv, 1953. It is presented here for comparative purposes.}
*Page=/77
*{(1)} **kí:***txt="ki:"*eng="this" **tsítsho***txt="jítsu"*eng="fox"
tehotstikawhenhákye'*txt="tehotstikawhʌhákye?"*eng="he-travel-along"
thontayawénhstsi*txt="thutayawʌ́hsji"*eng="suddenly" **yahà:rawe'***txt="yahà:lawe?"*eng="there-he-arrived" **tsi***txt=""*eng="" **nón:***txt="jinú:"*eng="place-where"
nikaya'kyón:ni*txt="nikaya?kyú:ni"*eng="place-she-lie down"
kohsá:tens*txt="kohsá:tʌs"*eng="horse" **yó:ta's***txt="yó:ta?s"*eng="she-sleeps"
*{This fox was travelling along, when suddenly he arrived at a place where a horse was lying down, horse sleeping.}

Ainsi la chaine en (12a) se verra attribuée la valeur en (12b) pour la propriété *Seg,* où le caractère ¦ sépare les morphèmes. Même si la segmentation s'applique à la forme lexicale, elle est effectuée en contexte en cliquant sur le mot dans l'interface Web de *Sato*. Dans cette édition, on affiche la forme segmentée, valeur de la propriété *Seg*, plutôt que la forme originale. On verra ainsi le texte s'enrichir au fur et à mesure qu'on procède à la segmentation des mots.

(12) APPLICATION DE *Seg* (*Sato*)
a. *tehotstikawhenhákye'* b. *te¦ho¦t¦stikawh¦en¦hákye¦* 'il errait'

Cette version segmentée sera ensuite exportée dans un format semblable à celui présenté au tableau 1. Le caractère ¦ étant un caractère séparateur dans l'alphabet *ka* que nous avons défini, la soumission à *Sato* de cette nouvelle version du corpus aura pour effet de générer un lexique de morphèmes qui pourront être catégorisés. Mais, avant d'être soumise à *Sato*, une opération supplémentaire sera effectuée sur le corpus au moyen d'expressions rationnelles évaluées par le langage Perl appelé depuis l'interface Web de *Sato*. Cette opération consiste à générer une forme non accentuée des morphèmes, tout en conservant la chaine originale dans la propriété *Acc* (Accent), tel qu'illustré en (13) :

(13) CHAINE SANS ACCENT (*Sato*)
*te*Acc="te"¦ho*Acc="ho"¦t*Acc="t"¦stikawh*Acc="stikawh"¦en*Acc="en"¦*
*hakye*Acc="hákye'"*txt="tehotstikawhʌhákye?"*eng="he-travel-along"*

Il en résulte une annotation lourde, mais l'interface Web n'affichera toutefois que la forme accentuée. Toute l'information peut être affichée dans une fenêtre séparée sous l'action d'un clic sur le morphème. L'annotation linguistique du corpus se poursuivra par l'ajout de propriétés sur les morphèmes. La catégorisation pourra être multiple, en particulier, catégorisation directe à l'écran inspirée de l'annotation de Bonvillain et application de dictionnaires de catégories générés depuis *ivi/Vinci*. Le développement d'automates de segmentation en morphèmes et de catégorisation, que nous entreprenons maintenant, pourra donc être comparée à l'annotation manuelle pour être ensuite appliquée à d'autres corpus[5].

[5] Il existe maintenant plus de consensus sur l'identification des frontières de mots dans les textes écrits. Conséquemment, le découpage de mots tel qu'il apparait dans la version publiée du texte *The Bear and the Fox* n'est pas le même qu'il ne serait aujourd'hui.

7. La catégorisation des morphèmes : résultats préliminaires.

La constitution d'un lexique de morphèmes nous a permis d'entrer dans *Sato* la catégorisation de Bonvillain (propriété *BV*). On a choisi d'utiliser une propriété lexicale cumulant les valeurs en contexte, s'il y a lieu, afin de comparer cette catégorisation avec celle générée par *ivi/Vinci.* On aurait pu s'attendre à ce que la catégorisation de Bonvillain soit exhaustive. Cette catégorisation est incomplète puisqu'elle fait appel à des étiquettes sémantiques pour les verbes, les noms et certains déterminants. Cette catégorisation ne fournit pas les catégories grammaticales, qui permettraient d'identifier, en tant que classes, les préfixes, les suffixes, les noms, les verbes, etc. C'est à ce niveau le plus simple que nous avons choisi de tester la capacité de la génération à compléter le balisage manuel en produisant un micro-dictionnaire que Sato appliquera sur le lexique des morphèmes compilé à partir des six premières phrases du récit *The Bear and the Fox* (propriété *Mvinci)*. Nous avons défini une série de règles de génération dans *ivi/Vinci* pour produire des étiquettes simples : PREF (préfixe verbale, à l'exclusion des préfixes pronominaux) ; PPREF (préfixe pronominal) ; V (verbe) ; SUFF (suffixe) ; N (nom). Dans la figure 2, nous comparons les résultats de cette double catégorisation sur les morphèmes reconnus par le micro-dictionnaire d'*ivi/Vinci*. La colonne *Ana* résume l'analyse. Le micro-dictionnaire généré par *ivi/Vinci* a reconnu 27 morphèmes sur 70. Il a attribué la bonne catégorie grammaticale (*V-ok*) dans plus de la moitié des cas (15 cas ; 56%) ; dans quatre autres cas (15%), *ivi/Vinci* propose une bonne analyse, mais avec ambigüité (*V-amb*), puisque le système propose deux possibilités. Il reste que dans huit cas (30%), une mauvaise analyse résulte de l'application du micro-dictionnaire (*V-bad*). Le tableau 2 présente les morphèmes identifiés comme *V-amb et V-bad.*

Figure 2 : Résultats comparés de la catégorisation

Totalfreq	BV	Mvinci	Ana	Form
1		(PPREF,PREF)	V-amb	sa
3	TRNS	(PPREF,PREF)	V-amb	y
1	CIS	(PPREF,SUFF)	V-bad	k
2	1s	(PPREF,SUFF)	V-amb	e
7	SRL / ITR	(PPREF,SUFF)	V-amb	s
1	3sM	SUFF	V-bad	h
3	PAR / SP	SUFF	V-bad	n
1	PRF	PREF	V-bad	en
1	EMP	PPREF	V-bad	ken
4	PAR	PPREF	V-bad	ni
3	say	PPREF	V-bad	ron
1	TRNS	PPREF	V-bad	ye

L'examen détaillé de ces cas met en évidence les causes de ces erreurs, qu'on peut repartir entre les facteurs suivants :

- Le micro-lexique comprend des fois une forme, mais non pas avec la valeur retrouvée dans corpus. Ainsi, le micro-lexique ne comprend pas la valeur perfective (PRF) pour la forme *en-*, mais seulement la valeur du futur qui se réalise comme préfixe, d'où l'erreur.
- Dans d'autres cas, un morphème du corpus prend une variante inconnue du micro-dictionnaire à cause de facteurs contextuels. Ainsi, bien que le micro-dictionnaire contienne *ho* (variante du préfixe pronominal *ro* en position non-initiale), et également le suffixe *h* (qui marque le perfectif), il ne comprend pas la valeur pronominale pour *h* quand ce dernier se manifeste comme variante de *ro* entre l'aoriste *a* et le préfixe réfléchi *at*.
- Finalement, dans un cas (CIS, troisième ligne), le micro-dictionnaire se trompe à cause d'une différence dialectale. Dans le dialecte utilisé pour formuler les règles de génération,

le cislocatif (CIS) prend la forme *ti-* ou une variante, mais dans le dialecte dans lequel le texte est transcrit, le cislocatif prend la forme *k-*, que le micro-dictionnaire associe à un préfixe pronominal ou à un suffixe. On voit donc qu'il sera nécessaire de paramétrer les règles de génération selon le dialecte des textes à traiter. La catégorisation lexicale opérée par *ivi/Vinci* amène nécessairement plusieurs catégories grammaticales sur certains morphèmes. C'est le propre d'un dictionnaire de rassembler toutes les catégories et les significations instanciées dans la multitude des contextes. Il revient à l'analyse en contexte de déterminer laquelle des catégories est instanciée, en tenant compte en particulier des contraintes syntaxiques s'appliquant à la séquence de morphèmes. Par exemple, quand le dictionnaire nous indique que le morphème pourrait se trouver en position préfixale, on peut enlever la catégorie PREF pour une occurrence donnée, si celle-ci suit un verbe.

Dans le logiciel SATO, on peut élaborer un scénario de commandes qui appliquera une série de règles contextuelles de désambigüisation, comme l'illustre la figure 3. Dans cet exemple, les lignes qui débutent par * sont des commentaires.

Figure 3 : Exemple de scénario de désambigüisation en contexte

```
* Scénario de résolution en contexte des ambigüités grammaticales illustrant la levée de l'ambiguïté de « en » dans la première phrase.
* On a d'abord corrigé la catégorisation de « en » en lui ajoutant la valeur manquante, donnant une une double catégorisation : (PREF,PRF). Et on projette la catégorisation lexicale (propriété Mvinci) sur une propriété sur les occurrences (Syn).

property Define  Syn inheritance MVinci for text

* On définit la borne des contextes comme étant le paragraphe (qui correspond ici à la phrase).
context characterize bounds = delimited $*Edition=par excluded $*Edition=par included

* On applique un patron contextuel pour enlever la catégorie PREF aux occurrences à la droite du verbe. Le patron contextuel est une suite de deux filtres s'appliquant chacun à une position.

$*Syn=V*.10 trouve toutes les chaines ($) dont la propriété Syn=V ; *.10 indique que le second filtre s'applique aux morphèmes suivants jusqu'à une distance de 10 ; $*Syn=PREF*Syn:-PREF*+ trouve les occurrences suivant le verbe dont Syn=PREF et leurs enlève la catégorie PREF (*Syn:-PREF)
context apply $*Syn=V*.10 $*Syn=PREF*Syn:-PREF
```

8. Conclusions provisoires/prospectives

Même si le premier test de notre prototype expérimental ne couvre qu'une fraction du texte, on peut déjà dégager certaines conclusions qui pourront guider la suite de la recherche.

Bien que la génération offre la possibilité de combler des lacunes de l'annotation de Bonvillain, il nous faudra fournir à *ivi/Vinci* un lexique plus étendu et une plus grande gamme de variantes orthographiques pour chaque forme canonique du morphème. On devra aussi compléter notre ébauche de scénario de désambigüisation en contexte.

On doit aussi constater la nécessité d'établir un système de catégories qui puisse rendre compte de la complexité morphologique des langues autochtones des Amériques, dont le kanyen'kéha est une parmi plusieurs autres. Les efforts du *UniMorph Project* (Sylak-Glassman 2016) pour établir un schéma universel de traits morphologiques basé sur une approche interlinguistique, est un cadre de référence que nous avons commencé à confronter aux langues autochtones des Amériques. Il nous apparait déjà que des traits devront être ajoutés au schéma pour tenir compte

de ces langues et unifier les descriptions qui ont été utilisées dans les recherches linguistiques déjà réalisées ou en cours de réalisation.

Références.

Baines, D. (2018). An Overview of FieldWorks and Related Programs for Collaborative Lexicography and Publishing Online or as a Mobile App. In (J. Čibej, V. Gorjanc, I. Kosem, S. Krek, eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, 953-958. https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2927-1-10-20180820.pdf .

Barrie, M. et S. Jung (2020). The Northern Iroquoian nominalizer and lexical categories. *Canadian Journal of Linguistics/Revue canadienne de linguistique,* vol. 65(1): 1–24.

Boyland, J. T. (1996). Conditional Attribute Grammars. *ACM Transactions on Programming Languages and Systems*, vol. 18(1): 73-108.

Bonvillain, N. et B. Francis. (1970). *The Bear and the Fox*. In M.A. Mithun et A. Woodbury, *IJAL Native American Text Series,* vol. 4: 79-95.

Daoust, F. Logiciel SATO, version 4.4 (rev. 2018) et 4.5 (2000) http://sato.ato.uqam.ca/

Chomsky, N. (1959). On certain formal properties of grammars. *Information & Control*, vol. 2(2): 137-167.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Foster, M.K. (1985). The language of tense, mood, and aspect in Northern Iroquoian descriptions. *International Journal of American Linguistics*, vol. 51: 403-405.

Foster, M.K. (1986). Updating the terminology of tense, mood, and aspect in Northern Iroquoian descriptions. *International Journal of American Linguistics*, vol. 52: 65-72.

Green J. (2018) Kanyen'kéha : langue mohawk. *Encyclopédie canadienne*. https://www.thecanadianencyclopedia.ca/fr/article/kanyenkeha-langue-mohawk (consulté le 27 janvier 2020).

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., Hulden, M. (2018). UniMorph 2.0: Universal Morphology**.** In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation), (ELRA):* 1868-1873. https://www.aclweb.org/anthology/L18-1293.pdf

Knuth, D. (1968). Semantics of Context-Free Languages. *Mathematical Systems Theory*, vol. 2(2): 127-145.

Lessard, G., Levison, M. VINCI Laboratory, http://research.cs.queensu.ca/CompLing/

The Mohawk Language Steering Committee. (1993). *The Mohawk Language Standardization Project: Conference Report* August 9-10. https://kanienkeha.net/the-mohawk-language-standardisation-project/ (consulté le 20 janvier 2020).

Littell, P., Kuhn, R., Hall, A., Kazantseva, A., Pine, A., & Cox, C. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. In J. L. Klavans (Ed.), *Proceedings of the 27th International Conference on Computational Linguistics* : 2620–2632). https://www.aclweb.org/anthology/papers/C/C18/C18-1222/ (consulté le 20 janvier 2020).

Mithun, M. (2006). Iroquoian Languages, dans *Encyclopedia of Language & Linguistics,* vol .6: 31-34.

Sylak-Glassman, J. (2016). *The Composition and Use of the Universal Morphological Feature Schema* (UniMorph Schema), Working Draft, vol. 2. Center for Language and Speech Processing, Johns Hopkins.