Traitement de la morphologie du finnois par transducteurs à nombre fini d'états

Marie CALBERG

OSTERLITS

marie.calberg@wanadoo.fr

Mots-clefs – Keywords

Dérivation – suffixe – finnois – tranducteurs à nombre fini d'états

Derivation – suffix – finnish – finite state tranducers

Résumé - Abstract

Cette étude présente un modèle pour le traitement de la morphologie du finnois. Ce modèle est fondé sur des transducteurs à nombre fini d'états. L'approche utilise une façon originale d'organiser les données et de générer dynamiquement une structure sémantique à partir d'une analyse morphologique. L'approche est linguistiquement validée par une étude des suffixes de dérivation verbale en finnois.

This study presents a model for Finnish morphology processing. This model is based on finite-state machines. The approach uses an original way to structure the data and dynamically generates semantic information from the morphology analysis. The approach is linguistically validated by a study of Finnish verbal suffixes.

1 Introduction

Du point de vue de leur structure, les mots finnois peuvent être des mots racines, dérivés ou composés. Le noyau du vocabulaire est formé par des mots racines : ces mots ne sont pas décomposables et leur thème se réduit au morphème radical. À côté de la composition, la dérivation est le procédé le plus important pour la formation de mots nouveaux. Notre hypothèse de départ est de voir s'il existe une corrélation entre l'emploi des suffixes de dérivation verbale et la nature d'un texte.

L'article présente un rapide état de l'art. Nous proposons ensuite un système reprenant la notion de morphologie à deux niveaux enrichie de processus de structuration et d'héritage au sein des éléments constitutifs du lexique.

2 Travaux antérieurs en traitement automatique de la morphologie du finnois

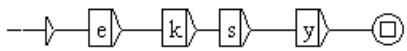
Ces dernières années, de nombreux travaux concernant le traitement automatique de la morphologie des langues naturelles ont été développés. Nous mentionnons ici trois sources d'inspiration. K. Koskenniemi (1983) propose un modèle à deux niveaux pour le traitement de la morphologie. Le premier niveau correspond aux unités morphologiques de base tandis que le deuxième niveau code les variations morphologiques dues à la mise en correspondance d'unités de base au sein d'unités de plus haut niveau. D. Clémenceau (1996) a également proposé un modèle fondé sur un système de transducteurs à nombre fini d'états pour traiter la morphologie. R. Evans et G. Gazdar (1996) ont proposé un modèle riche permettant de structurer hiérarchiquement les unités linguistiques de sorte que des informations linguistiques peuvent être partagées par plusieurs unités. Le lexique ainsi développé est donc moins redondant et plus cohérent.

3 Un modèle pour traiter la morphologie du finnois

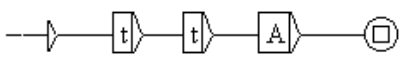
Le modèle que nous nous avons développé pour analyser la morphologie du finnois est fondé sur un ensemble de transducteurs nombre fini d'états. Il est inspiré de différents travaux, entre autres ceux de Koskeniemmi (morphologie à deux niveaux, 1986), de Clémenceau (morphologie à nombre fini d'états, 1993), de Gazdar (organisation hiérarchique du lexique, cf. DATR, 1996).

3.1 Les données de base (morphologie de premier niveau)

Les différents morphèmes sont d'abord décomposés dans des graphes indépendants. Voici par exemple le graphe lexical qui correspond à la racine *eksy* (*perdre*):



Le graphe suivant correspond au dérivé causatif :

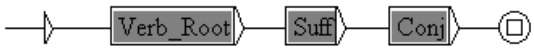


Et le graphe suivant correspondant à l'infinitif :

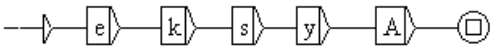
Chaque graphe de base est typé : les racines lexicales sont typées en `Verb_Root`, les suffixes de dérivation en `Suff` et les terminaisons verbales en `Conj`. La combinaison de graphes

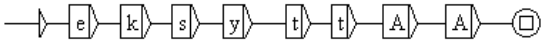
indépendants est codée en utilisant un métagraphe lexical abstrait permettant d'appeler dynamiquement les graphes lexicaux. Par exemple, le graphe suivant :

correspond à un verbe simple, tandis que le graphe ci-dessous :



correspond à un verbe suffixé. La correspondance entre le nom des « boîtes » dans les graphes ci-dessus et les types des graphes élémentaires est évidente (ce point est modelisé par les boîtes grises, qui renvoient aux sous-graphes dont elles portent les noms). Nous obtenons ainsi les graphes lexicaux suivants :

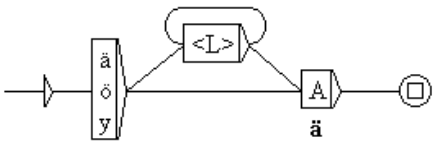




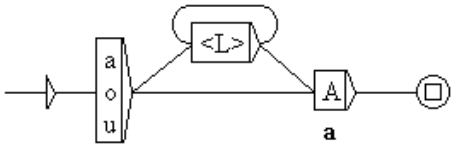
qui correspondent respectivement à la forme simple et dérivée du verbe dont la racine est *eksy*. Nous examinons dans la section suivante les règles de linéarisation permettant de produire des unités lexicales valides. Un ensemble de règles de linéarisation permettent en outre de contraindre l'insertion des voyelles entre la racine et le suffixe (cf. `tunn-i-sta-a <- tuntea`). Enfin, des règles permettent de modifier dynamiquement certains cas de mise en relation de racines avec des suffixes (phénomènes d'assimilation). Clémenceau (1996) a proposé un modèle permettant d'implanter ce type de propriétés linguistiques à l'aide des machines à nombre fini d'états. Nous nous inspirons ici de l'analyse proposée par Clémenceau.

3.2 Problèmes de linéarisation (morphologie à deux niveaux)

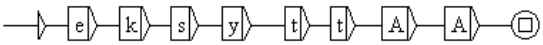
L'approche que nous avons proposée n'est pas complète puisque ni **eksya* ni **eksyttA* ne correspondent aux formes lexicales valides en finnois. La linéarisation est un mécanisme prévu pour produire les formes de surface valides, c'est-à-dire eksyä de **eksya* du fait que le radical de verbe contient une voyelle antérieure (y et pas u). Les problèmes de linéarisation sont résolus en utilisant des graphes spécifiques qui sont appliqués après l'étape précédente. Les lettres majuscules dans les graphes lexicaux correspondent à des variables qui doivent être repérées à partir du contexte. Le graphe suivant code ce genre d'information :



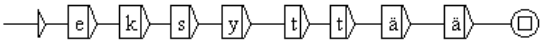
Le graphe indique le fait que la variable (la lettre majuscule A) doit être réalisée par un ä si et seulement si elle est précédée par une voyelle antérieure ä, ö ou y. Le <L> correspond à n'importe quelle lettre de l'alphabet. Le graphe correspondant aux voyelles postérieures est le suivant :



Enfin, un graphe par défaut s'applique si le radical de verbe contient des voyelles neutres. Ces graphes de linéarisation sont appliqués après les graphes lexicaux précédents. Par exemple, à partir du graphe ci-dessous :



on peut obtenir le résultat suivant :

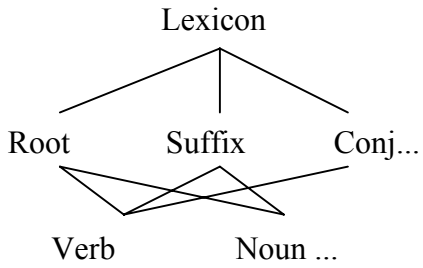


ce qui correspond à une forme de surface valide.

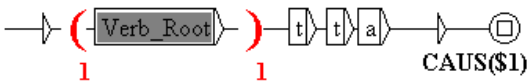
3.3 Les propriétés sémantiques

La décomposition de morphèmes lexicaux en unités de base permet de leur associer une information sémantique. Ceci permet de générer dynamiquement une structure sémantique pour les unités lexicales complexes, par analyse compositionnelle. Ceci est particulièrement important en finnois du fait qu'il s'agit d'une langue agglutinante.

Nous associons alors des informations sémantiques aux unités morphologiques de base. Ces informations peuvent être attachées à différents niveaux de détail, puisque les unités morphologiques sont stockées dans un treillis inspiré de DATR (Gazdar et Evans, 1996) :



Ce schéma est particulièrement important pour le codage de l'information sémantique. Par exemple, le transducteur suivant :



permet de calculer la forme sémantique du verbe comme *eksyttää*. La boîte `Verb_Root` correspond à la racine *eksy* de sorte que le transducteur produit la forme suivante :

```
eksyttää -> le CAUS (eksy)
```

La généralisation de ce type de codage est particulièrement adaptée pour la morphologie du finnois. Il est naturellement impossible de stocker l'ensemble des formes linguistiques de surface dans un dictionnaire pour une langue agglutinante comme le finnois; le lexique serait infini. Une approche générative permet un traitement plus économique et plus approprié à la morphologie du finnois.

3.4 Algorithme

L'algorithme proposé est alors très simple. Il est tout d'abord nécessaire d'identifier les formes ambiguës, c'est-à-dire les formes linguistiques qui se terminent comme un dérivé mais qui n'en sont pas. Ces formes ambiguës sont enregistrées dans un dictionnaire spécifique. L'analyse des autres formes est alors effectuée par l'application du lexique sous forme de graphes à nombre fini d'états. L'algorithme suivant est appliqué :

1. Isoler les formes ambiguës
2. Appliquer le lexique sous forme de transducteurs sur les formes non ambiguës
3. Générer la représentation sémantique attachée aux formes de surface analysées

Cet algorithme a été appliqué sur un corpus représentatif pour valider les hypothèses linguistiques de départ.

3.5 Implémentation

Les graphes ci-dessus ont été dessinés en utilisant le module graphique d'INTEX/UNITEX (Silberztein, 1993). Cependant, l'application récursive des graphes morphologiques indépendants n'a pas encore été vraiment implanté dans cette plateforme. Par exemple, il n'est pas possible de décomposer dynamiquement des entrées complexes pour produire une analyse morpho-sémantique.

Nous avons développé une première implantation, en utilisant des scripts Perl. Cette implantation est fondée sur un ensemble d'expressions régulières codant l'analyse morphologique détaillée ci-dessus. Ce travail fait partie d'un projet de plus grande ampleur prévu pour fournir une analyse complète du finnois en utilisant les transducteurs à nombre fini d'états (cf. dernière section : perspectives).

4 Conclusion

Nous avons présenté dans cet article un modèle pour traiter la morphologie de langues agglutinantes telles que le finnois. L'approche est fondée sur la manipulation d'unités de base codant des informations morphologiques et sémantiques. Des règles constructivistes permettent de calculer une représentation sémantique pour les unités complexes et de contrôler les phénomènes d'assimilation de voyelles. Ce travail est développé avec des linguistes et permettra la validation d'hypothèses linguistiques sur la structure du finnois. Le travail a été validé sur un large corpus textes en finnois. (Calberg, 2002).

5 Perspectives : des données à grande échelle pour le traitement du finnois

L'expérience décrite dans cet article fait partie d'un projet de plus grande ampleur prévu pour fournir des ressources à grande échelle pour le traitement du finnois. La première étape consiste à compléter les ressources existantes. Nous envisageons une collaboration pour développer un dictionnaire et un étiqueteur pour cette langue. Le résultat devrait être utilisable dans la plateforme d'INTEX/UNITEX (Silberztein, 1993). L'approche constructiviste et générativiste décrite dans cet article sera, autant que peut se faire, respectée pour fournir une description riche de cette langue. Enfin, un corpus représentatif plus important doit être établi pour le finnois, comme il en a été discuté ci-dessus. Ce travail sera développé par des informaticiens et des linguistes pour établir une approche réellement multidisciplinaire permettant d'aborder la linguistique de corpus en finnois dans ses différentes dimensions.

Remerciements

Nous remercions Thierry Poibeau, M.M.Jocelyne Fernandez-Vest, André Salem et Serge Fleury pour leur soutien à cette participation et les relecteurs de RECITAL pour leurs remarques qui nous ont permis d'améliorer cet article.

References

Blåberg O. (1994) *The Ment Model. Complex States in Finite State Morphology*. PhD dissertation. Dept. of Linguistics. Uppsala, Sweden.

Calberg M. (2002) « Les suffixes de dérivation verbale du finnois et leurs équivalents français - approche en Traitement Automatique des Langues ». *6e colloque contrastif français-finnois*, Helsinki.

Clémenceau D. (1996) "Finite-state morphology processing : inflections and derivation in a single framework using dictionaries and rules". In Roche E. and Schabes E. (eds.) *Finite-state language processing*, MIT Press, Cambridge.

Dubois J., Giacomo M., Guespin L., Marcellesi C. et Mevel J-P. (1994). *Dictionnaire de linguistique et des sciences du langage,* Larousse, Paris.

Evans R. and Gazdar G (1996) "DATR: a language for lexical knowledge representation". *Computational Linguistics*, 22(2), pp. 797-815.

Goldsmith (2000) "Unsupervised Learning of the Morphology of a Natural Language". *Computational Linguistics* 27(2), pp. 153-198.

Habert B., Nazarenko A. et Salem. A. (1997). *Les linguistiques de corpus*. Armand Colin Paris.

Koskenniemi K. (1983) *Two level morphology : a general computational model for word form recognition and production.* University of Helsinki.

Silberztein M. (1993) *Dictionnaires électroniques et analyse automatique de textes, le système Intex*. Masson, Paris.