Visualisation en arbres pour assister l'étude des sons du nisvai

Jocelyn Aznar

Aix-Marseille Université, EHESS, CREDO

LIFT 2019

Linguistique informatique, formelle & de terrain

Orléans
28 et 29 novembre

Objectifs

La représentation en arbres des morphèmes d'un corpus accompagne l'analyse de la distribution des sons et phonèmes. Le retour visuel fourni par les arbres facilite la conception d'hypothèses sur les processus phonologiques opérant dans la langue et permet d'évaluer la cohérence de la transcription.

Introduction

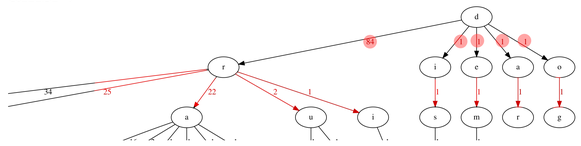
La visualisation en arbres de la transcription a été utilisée lors de l'étude de la phonologie du nisvai, une langue orale parlée par environ 200 locuteurs dans le sud-est de Malekula, au Vanuatu.
La transcription du nisvai a servi de base pour proposer une écriture à la communauté nisvaie et a facilité l'étude des textes oraux. Elle a été réalisée avec les locuteurs lors d'entretiens sur la pratique de la langue. Qu'elle soit phonétique ou phonologique, la transcription a été une étape durant laquelle des hypothèses ont été formulées sur les structures de la langue et les variations observées.
Pour produire ces arbres, un script Python a extrait et segmenté les morphèmes lexicaux en caractères. Le résultat a été formalisé dans le langage DOT, et la représentation visuelle a été assurée grâce au module Graphviz (Ellson et al., 2004).

Le corpus de données

Le corpus étudié réunit 34 textes nisvai oraux d'une durée totale de 174 minutes. L'annotation des textes narratifs nisvais a été réalisée avec le logicel ELAN (Brugman & Russel, 2004), pour être ensuite intégrée au sein d'une base de données. Ces textes totalisent 86495 caractères répartis en 11346 morphèmes lexicaux et 20206 morphèmes grammaticaux. Seuls les morphèmes lexicaux ont été retenus pour produire les arbres présentés ici.

Arbre et forêt des occurrences

Un arbre pondéré par les occurrences se présente ainsi :



Parmi les 88 occurrences de la lettre d à l'initial, 84 sont suivies par la lettre r, alors qu'il n'y en a que 4 pour les lettres i, e, a, o. Chacune de ces occurrences est prolongée si une branche plus développée existe. Chaque arbre de la forêt correspond à une lettre possible à l'initial d'un morphème.

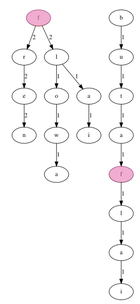
La forêt de transcriptions du nisvai

Lettres	r	t	p	q	s	d	f	g	h	j	k	l	m	w	c	v	b	n	a	e	y	u	i	o
Arbres	18	17	10	11	17	13	2	17	18	1	13	19	20	12	15	18	19	21	21	20	15	19	19	21
Branches	327	144	108	54	205	63	4	113	266	1	121	254	221	45	154	212	251	382	688	202	100	305	260	221

Totaux arbres : **Voyelles :** 115 **Consonnes :** 261 **Somme :** 376
Totaux branches : **Voyelles :** 1776 **Consonnes :** 2925 **Somme :** 4701

Observer les sons empruntés

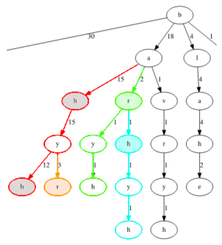
Les sons nisvais empruntés sont situés dans des arbres moins développés que les arbres des autres sons :



Il n'existe pour la lettre *f* que deux arbres. Le tableau précédent montre que les autres sons, à l'exception de *j*, autre emprunt, sont situés au sein d'une forêt plus importante.

Variations lors de la transcription

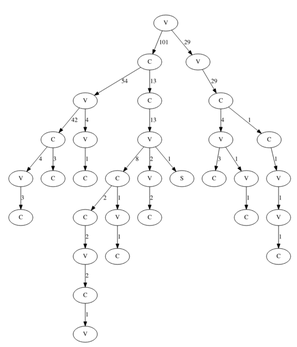
La représentation en arbre facilite l'observation des variations graphiques. Les couleurs, ajoutées pour l'exemple, font ressortir les quatre branches du morphème /bahyh/ : "tronc".



Nous voyons ainsi les variations de transcription entre la lettre r : /ɾ/ et la lettre h : /ɣ/.

Étudier les caractères par classe

La représentation en classes de caractères permet de caractériser le comportement des segments phonologiques.



Cette arbre montre que le nombre de branches, lorsqu'une voyelle est à l'initial d'un morphème lexical, est limité.

Conclusion & Perspectives

La synthèse que propose la visualisation en arbres pondérés a aidé à répérer les régularités et les anomalies parmi les transcriptions du corpus. Elle a contribué à analyser les sons du nisvai comme phonèmes ou allophones et a permis d'identifier les structures les plus courantes ou au contraire les hapax.
Améliorer l'utilisabilité de la production des arbres est un enjeu important, car l'un des intérêts de la visualisation est de faciliter l'interrogation des données. Trois pistes sont envisagées dans cette perspective :

- Relier les arbres aux données primaires.
- Proposer la recherche par groupes de caractères.
- Permettre de définir des contextes aux requêtes afin de limiter le nombre de résultats retournés.

Références

Brugman, H. & Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 2065–2068).

Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., & Woodhull, G. (2004). Graphviz and dynagraph — static and dynamic graph drawing tools. In M. Jünger & P. Mutzel (Eds.), *Graph Drawing Software* (pp. 127–148). Springer Berlin Heidelberg.

Remerciements

Je souhaite particulièrement remercier la famille Gelu et les personnes de la communauté nisvaie pour leur hospitalité et leur patience.

Courriel

contact@jocelynaznar.eu