Iles d Imesli, 9, pp.183-189

Un outil d'élaboration de dictionnaire multimédia pour la langue tamazigh

Philippe MARTIN
Professeur Émérite
LLF UMR7110, UFRL
Université Paris Diderot Sorbonne Paris Cité

Agzul

Deg uḍris-agi, nebɣa ad nesken yiwen wallal swacu nezmer ad d-nexdem amawal umḍin deg tmaziɣt. Alla-agi, yezmer ad d- yesuffeɣ amawal imawi s tmaziɣt, ad d-yessekcem akk tantaliwin d tirawin yellan.

Abstract

A software tool devoted to the elaboration of a Tamazigh dictionary allows the incorporation of speech data as well as their acoustic segmental and melodic characteristics in addition to the classical monolingual dictionary pattern which includes examples of linguistic usage and their translation in different languages (Arabic, French, English).
This tool integrates the characteristics of popular spreadsheets such as Excel, as well as those of an advanced phonetic analysis software. It allows also an easy evolution of its characteristics, in order to correspond to the evolving users requirements appearing during the integration of the linguistic data of the dictionary.

Keywords: dictionnary, Tamazigh, speech data, lexical, lexiphonic

Les dictionnaires monolingues

Jusqu'à une époque récente, les dictionnaires utilisaient, pour illustrer les définitions de chaque entrée, des exemples transcrits en orthographe standard ou par une représentation phonétique. Le passage obligé par une médiation écrite est d'ailleurs caractéristique de la plupart des recherches en linguistique alors que leur objet consiste en fait de syllabes, de mots, de syntagmes, d'énoncés entiers ou même de tours de parole ou de conversations, objets prononcés et non écrits au départ. Il en résulte pour la recherche linguistique un biais, un parti pris pour les formes écrites de l'activité langagière au détriment de sa forme orale originelle. On aboutit ainsi à une linguistique de l'écrit, ne rendant qu'imparfaitement compte des activités linguistiques réelles des sujets parlants.
De même, les dictionnaires, monolingues ou non, ne donnent comme entrée que la version écrite des mots, le plus souvent sous une forme de référence telle que l'infinitif des verbes ou la forme masculin singulier des noms, etc. L'apparition de dictionnaires informatisés et de dictionnaires en ligne a

certes permis d'ajouter à la forme écrite des entrées une version pouvant être oralisée, mais ces formes ne correspondent en général qu'au seul mot correspondant à l'entrée.
Pour remédier à ces limitations, on présente dans ce chapitre un outil informatique permettant de réaliser un dictionnaire de l'oral ayant les mêmes caractéristiques que les dictionnaires de l'écrit, en donnant non pas une simple oralisation du mot sélectionné à l'entrée, mais également des exemples d'utilisation de ce mot dans la parole réelle. On obtient ainsi l'exact équivalent oral du dictionnaire conventionnel orthographique, pour lequel les définitions d'un mot, paraphrases de collections d'exemples, sont illustrées par des exemples dans des contextes variés permettant d'apprécier les différentes significations possibles d'une même entrée lexicale.

Un tel dictionnaire de l'oral serait également le contrepoint des grandes bases de données utilisées en traitement automatique du langage, en particulier en traduction automatique, et servirait aux recherches actuelles pour obtenir une grammaire utilisable par les algorithmes de reconnaissance de la parole spontanée.

Un outil : WinPitch

L'élaboration d'un dictionnaire de l'oral implique l'utilisation d'outils informatiques dédiés et performants. Un tel outil, WinPitch présentés succinctement ici, a été continuellement développé depuis plus de 20 ans pour permettre la constitution de grands corpus oraux de parole spontanée, en particulier dans le cadre du projet européen C-Oral-Rom (2005). En plus des nombreuses fonctions d'analyse acoustique de la parole, WinPitch permet l'indexation facile de données orales transcrites (ou à transcrire) afin de retrouver les différentes prononciations et usages dans différents contextes sémantiques et syntaxiques présents dans un ensemble d'enregistrements donné.
Trois opérations de base sont nécessaires pour la mise en œuvre d'un projet de dictionnaire monolingue de l'oral : 1) la transcription des enregistrements, 2) l'alignement texte-parole et 3) la mise en place d'un concordancier.

Transcription

Plusieurs logiciels de transcription de parole sont disponibles, tels que Praat (www.praat.org) ou Transcriber (trans.sourceforge.net). Outre les fonctions de base de la transcription, WinPitch (www.winpitch.com) possède de plus des fonctions de segmentation automatique rendant les opérations plus faciles et plus rapides. Ainsi une pré-segmentation opérant à partir des pauses identifiées dans le signal assigne automatiquement une transcription arbitraire, que l'opérateur n'a plus qu'à mettre à jour en écoutant le segment de parole correspondant (retrouvé par un simple clic), éventuellement à vitesse ralentie sans modification des caractéristiques segmentales (Fig. 1).

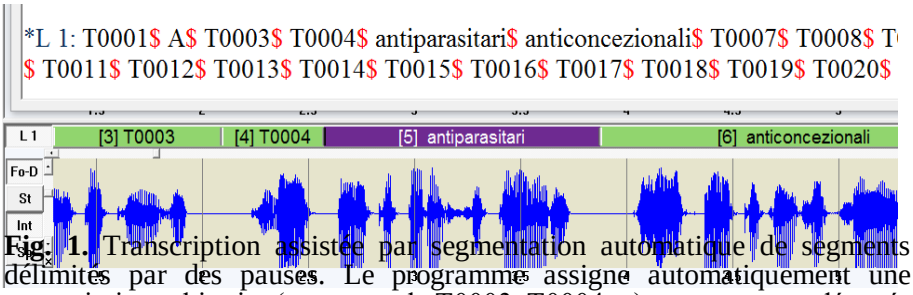

Fig. 1. Transcription assistée par segmentation automatique de segments délimités par des pauses. Le programme assigne automatiquement une transcription arbitraire (par exemple T0003, T0004…) aux segments détectés entre deux pauses, transcription que l'opérateur met à jour en écoutant les segments de parole correspondants.

Alignement texte-parole

Une autre caractéristique unique du logiciel WinPitch est l'alignement à la volée, permettant d'aligner un enregistrement dont la transcription est déjà disponible. Le principe de cette fonction est de permettre à un opérateur de cliquer sur tout élément de la transcription jugé pertinent (une syllabe, un mot, un syntagme…), texte affiché sur une fenêtre dédiée, tout en écoutant l'enregistrement à vitesse ralentie (jusqu'à 7 fois). Ce ralenti rend possible et facile la synchronisation du geste de positionnement de la souris avec la perception du son correspondant à l'élément du texte. L'opérateur positionne ainsi des balises insérées dans le texte et définissant automatiquement des postions temporelles dans l'enregistrement de parole (Fig. 2). En pratique, un ralenti de 30% à 50% est optimal pour réaliser cette opération à la volée, permettant de réaliser un alignement en deux ou trois fois le temps réel, tout en évitant les écueils de l'alignement forcé automatique, peu performant en parole bruitée. Des commandes ergonomiques à la souris (boutons gauche et droite, roulette) permet des corrections faciles en cas d'erreur (retour à la dernière balise, changement de ralenti, etc.). Cet alignement à la volée peut être sauvegardé dans divers formats (wp2, TextGrid, XML, etc.). On peut ensuite retrouver automatiquement le segment de parole correspondant à un mot ou un groupe de mots sélectionnés dans le texte à partir du soulignage des segments de texte souhaités.

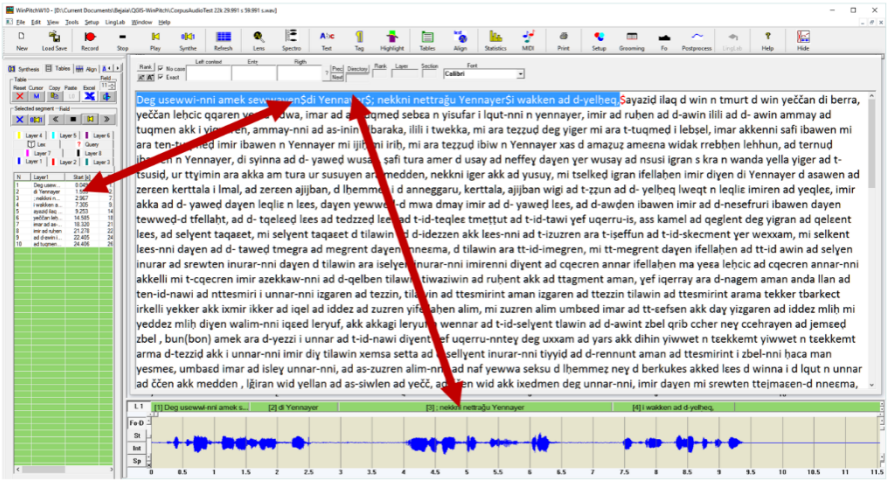

Fig. 2. Alignement à la volée avec parole ralentie. L'opérateur clique sur les mots du texte affiché au fur et à mesure de l'écoute à vitesse ralentie des segments de parole correspondants.

Concordancier

À partir d'un mot choisi par l'utilisateur, un concordancier intégré dans WinPitch retrouve toutes les occurrences de ce mot dans l'ensemble des transcriptions d'un même répertoire ainsi que leur contextes gauche et droit. En cliquant sur une occurrence affichée par le concordancier dans un tableau, le segment de parole correspondant et son analyse acoustique sont automatiquement affichés (Fig. 3).

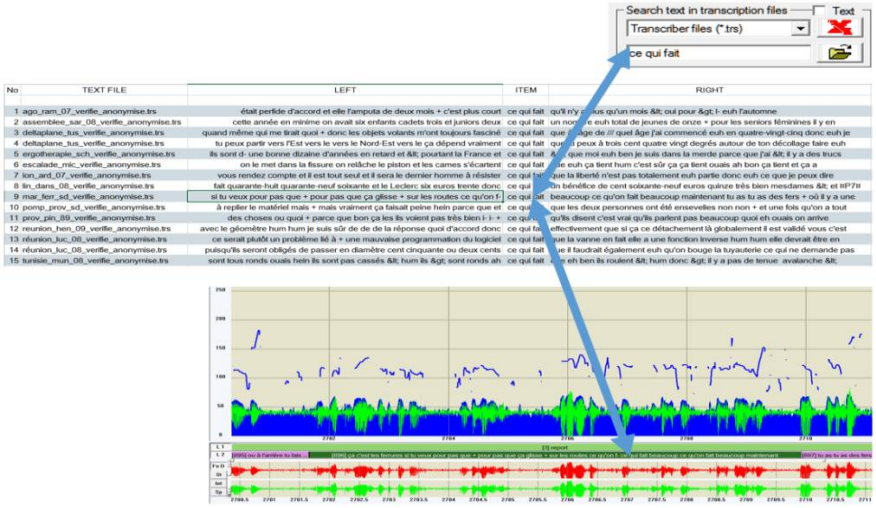

Fig. 3. Concordancier intégré. En cliquant sur une entrée du concordancier (ici implanté dans un tableur Excel®, le segment de parole est affiché automatiquement ainsi que l'analyse acoustique correspondante (courbes mélodique et d'intensité, spectrogramme).

Intégration dans un dictionnaire

Le concordancier intégré dans WinPitch (Fig. 3 et Fig. 4) donne automatiquement des listes d'exemples d'emploi d'une entrée du dictionnaire. De plus, la prononciation de ce mot est automatiquement analysée avec son contexte, ce qui est particulièrement important pour les études phonétiques.

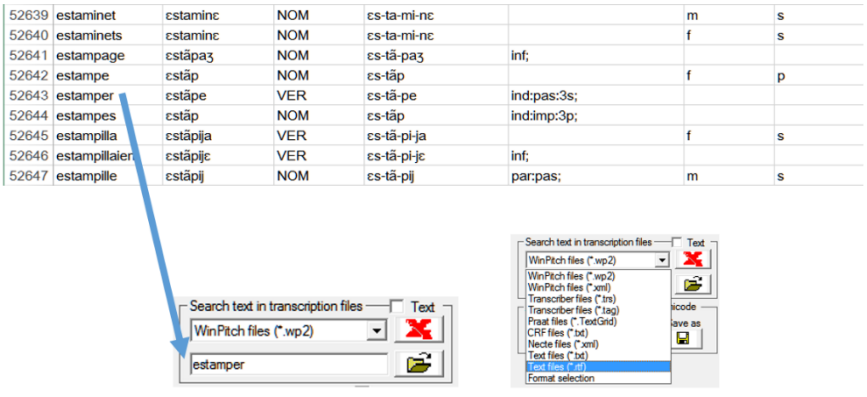

Fig. 4. Intégration dans un dictionnaire

Production automatique d'exemples

Le concordancier intégré dans WinPitch extrait automatiquement les exemples d'utilisation d'un mot sélectionné par l'utilisateur ainsi que leurs contextes, sous une forme non seulement orthographique mais également sonore (Fig. 5). Ces exemples sont donc prononcés dans leurs contextes respectifs, et non isolément comme dans les dictionnaires informatisés récents. On obtient ainsi directement des collections d'exemples pouvant être utilisés pour des recherches variées.

l'évolution quoi + mais nous ne sommes pas dans notre sujet voilà voilà voilà	alors	euh on a fait le tour des trucs là ouais < et moi il y avait < euh après
on a fait le tour des trucs là ouais < et moi il y avait < euh après >	alors	< voilà ah oui > alors on a fait musique chant pour l'art pla je vais
ouais < et moi il y avait < euh après > alors < voilà ah oui >	alors	on a fait musique chant pour l'art pla je vais vous dire que pour l'art pla
¤N3¤ > ouais oui c'est vrai ouais + je crois /// et en terme de théâtre	alors	pour le théâtre euh donc il y a la troupe de ¤T2¤ les troupes de théâtre de
en a parlé + que l'¤A2¤ ah non < l'¤A2¤ de ¤T2¤ > l'¤A2¤ donc euh oui +	alors	il y aurait danse de salon hip hop et danse africaine bon danse africaine ça
euh < et pourquoi > il y a un rapport avec le Ombres et Lumières	alors	non pour euh pour pour la partie musique ce qu'on a oublié de dire donc il y a
l'ADDAM 54 c'est euh une euh c'est une organisation l'ADDAM c'est ou une euh	alors	c'est une branche du conseil général euh ADDAM je sais pas exactement ce que ça
dit qu'on était quatre ah ouais il y a pas de monde il faut < recommencer	alors	> ¤P9¤ non mais on disait qu'il y avait pas beaucoup de monde parce que + la
groupe-là euh la thématique euh générale ce serait Reflets Ombres et Lumières	alors	donc c'est un peu < le fil rouge c'est une question > de Yohann ça c'est
une question > de Yohann ça c'est le fil rouge parce que < j'ai toujours	alors	> le fil rouge <du j'ai pas trouvé > le rapport moi je connais pas
aussi un petit spectacle itinérant pour les enfants des différentes écoles	alors	c'est ce qu'on disait tout à l'heure il y aurait euh donc euh ¤A3¤ pour les
il y aurait euh donc euh ¤A3¤ pour les tout-petits et ¤A4¤ pour les six-dix ans	alors	< tu as dit non > c'est pas ça ben si tu as dit ça tout à l'heure non
ben < oui c'est la version adulte de ¤A3¤ et c'est des pièces de théâtre ça	alors	c'est du euh + c'est du conte musical en fait ils racontent une histoire et au
dans les souks non ils nous ont laissés que trois quarts d'heure de temps	alors	euh c'était < chiant ça c'était > trop à la bourre hein le temps de
ils sont très sympathiques ah et puis le jeune il voulait qu'on te vende	alors	ouais ouais elle est belle la fille photo oh mince j'ai pas de photo de mes
> dans le désert à les à deux sur la mobylette et puis j'y vais je fonce	alors	comme il dit la M.B.K. dans le désert O.K. mais les autres non < d'accord
donc ben hum > ils traînent et là-bas ça doit être pareil ouais ça de- ou	alors	< les jeunes c'est n'ont peut-être > pas trop envie de travailler aussi
Internet tous les jours il y a des messes à telle heure et caetera hum ouais	alors	ça après euh avec Internet c'est vrai qu'on a on a accès à plein de trucs il y
tu as plein de de plastiques qui volent qui s'agrippent dans les oh oui <	alors	les > petits euh les petits feuillages qu'il y a par terre des < arbres à
oui oui < mais bon elle m'a dit > il en a on en a pas reparlé avec eux	alors	elle a dit nous on prend toujours des croissants quand on va chez ma fille à
quinze euros on a payé ah oui c'est rien elle paie ouais bah c'est bien bah	alors	cette fois-ci vraiment le manger impeccable oh oui ah bah on critique pas bah
j'aime pas quand c'est liquoreux ah il est bien sucré c'est sucré mais	alors	là il était excellent il avait fait du du hein Annie c'est sûrement Annie
puis après on a eu euh du fromage j'ai même pas mangé de fromage c'était trop	alors	il y a eu les entrées les petits toasts avant et tout non mais c'était bien
j'ai rencontré ¤P1¤ un Mexicain et euh il vendait des colliers tout et ¤P2¤	alors	là c'ét- totalement fou parce que ¤P2¤ c'est un Vosgien < et je demandais ah
et quand je referai un autre voyage j'en rencontrerai d'autres enfin ouais	alors	tu as l'intention de repartir ailleurs j'aimerais bien < déjà là >
belles hein et ils disent ils ont été vraiment brimés pendant la guerre	alors	ils veulent + < voilà c'est leur manière > à eux de de se venger un

Fig. 5. Production automatique d'exemples, ici avec le mot *alors*. En cliquant sur une ligne du tableau, WinPitch retrouve automatiquement le segment de parole correspondant au mot choisi avec son contexte.

Conclusion

Un dictionnaire monolingue oral, tel qu'élaboré grâce au logiciel WinPitch, permet l'intégration des différents dialectes, ainsi que les variantes orthographiques d'une même prononciation (transcriptions multicouches). Il n'est pas nécessaire d'effectuer le découpage de l'oral en mot isolé, chaque prononciation étant donnée dans et avec son contexte.
Du coup, la définition du mot comme espace entre deux blancs pourrait perdre sa circularité inhérente. En effet, contrairement au mot écrit, le mot oral n'existe que dans le cadre du groupe accentuel, séquence de syllabes contenant une seule syllabe accentuée (et correspondant généralement plusieurs mots de l'écrit, voir par ex. Martin, 2015). Il faut donc prendre acte de la disparition progressive actuelle du dictionnaire-livre pour utiliser les ressources informatiques disponibles, et ainsi passer de la lexicographie à la **lexicophonie**…

References

Cresti, Emanuela & Moneglia, Massimo (ed), 2005 : *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam, Benjamins, 304 p.

Martin, Philippe, 2015 : *The Structure of Spoken Language*, Cambridge, Cambridge University Press, 356 p.

WinPitch (2018) www.winpitch.com