Iles d Imesli, 7, 2015, pp. 151-161.

Présentation d'une démarche de valorisation et d'exploitation informatiques des corpus oraux. Le cas du logiciel Elan

Salem Djemai
Docteur en linguistique amazighe
Université de Tizi-Ouzou
djemai_salem@yahoo.fr

Agzul:
Di tezwart unadi-ya, yella-d wawal ɣef unbeddel ameqqran i d-yellan deg wallalen usekker asenselkam n wammuden i isemtawiyen ger imesli d tira n wammud. Deg weḥric-is amezwaru, nebder-d kra n tenfawin yesɛa Elan, diɣ neglem-d kra deg ugrudem-is, i nwala ɣur-s azal d ameqqran. Deg weḥric-is wis-sin, neṭṭef, d amedya, amek ara neg asnimek alɣasnay s Elan. Syin ɣer-s, nutlay-d ɣef tenfawin nniḍen, yellan deg Elan-CorpA, akka am usɣiwel n usnimek s umawal i d-yettakk a t-nefren. Di tagara-s, nesken-d kra iberdan unadi ɣef wayen yellan d isefka deg wammud i nesnimek di Elan d lexṣaṣ yellan deg-s, am uɣlluy ur d-ɣellint ara swa swa tilas imesli n wawalen d tira-nsen.

Abstract:
After the introduction, which emphasizes the upheaval that linguistics witnesses in data processing techniques corpus, which allow time synchronization between sound and annotations, the first part of this paper outlines the essential features that Elan is equipped and a brief presentation of the interface of this software. The second part, as an illustration, first shows how to use Elan for morphosyntactic annotation. Then it sets the additional features of ELAN-CORPA that simplify the interlinear in Elan for integrating a lexicon. Finally, it shows some options in Elan and discusses the limitations of the software, including the fact that it does not allow a precise location of the border segment of the sound signal on the note to which it refers that segment.

Longtemps, l'unique façon de diffusion des données recueillies par les chercheurs de terrain a été la publication sous forme de textes accompagnés de leur traduction et rarement de leurs enregistrements. Ces publications restent souvent confinées dans les bibliothèques et sont généralement difficiles à exploiter à cause de leur pauvreté en informations.

Actuellement, la linguistique connaît un bouleversement dans les techniques de valorisation et d'exploitation des corpus. En effet, des technologies récentes (Elan, Praat, Transcriber, etc.) permettent de numériser le son et d'avoir une synchronisation temporelle entre le signal sonore et une/des transcription(s) ainsi que des métadonnées. Selon le site du Consortium Corpus Oraux et Multimodaux :

Ces technologies peuvent non seulement permettre d'enrichir nos connaissances sur le contenu des corpus ou nous faire gagner du temps pour leur enrichissement préalable à l'analyse, mais également d'explorer des interfaces : ex. gestes-syntaxe, prosodie-sentiments, hésitations-stratégies d'argumentation, etc. En réalité les corpus numériques ouvrent de nouveaux champs de recherches grâces aux technologies émergentes et l'interdisciplinarité.[1]

Avec des transcriptions alignées sur le signal sonore, l'oral devient plus facile à analyser et l'accessibilité aux données devient très aisée. Cette démarche permet de faciliter la vérification et l'enrichissement des données, puisqu'on peut attribuer à chaque segment de multiples niveaux de transcription. Autrement dit, nous sommes à l'inverse d'une notation orthographique sur papier qui devient une forme immuable des données de base. Selon Abouda L. et Baude O. (2007 : 07) : « *Ici, la transcription n'est plus la vérité d'un chercheur (au mieux) ou d'un transcripteur, elle devient cumulative* »[2]. L'annotation est donc destinée à être une tâche collaborative, d'où la nécessité d'employer des conventions de transcription orientées vers l'interopérabilité.

Dans notre article, tel que le suggère le titre, nous essayerons de donner une présentation de l'un des outils techniques conçus pour l'annotation et l'exploitation des corpus linguistiques, en l'occurrence Elan. La première partie expose l'essentiel des fonctionnalités dont Elan est outillé et fait une brève présentation de son interface. La seconde partie, comme illustration, montre d'abord comment employer Elan pour une annotation morphosyntaxique. Ensuite, elle expose les fonctionnalités supplémentaires d'ELAN-CorpA, qui simplifient l'interalignement dans Elan. Enfin, elle montre quelques options de recherche dans Elan et évoque les limites de ce logiciel.

I. Présentation d'Elan

Elan (**E**UDICO **L**inguistique **An**notator) est un logiciel qui a été développé à l'Institut Max Planck de Psycholinguistique de Nimègue au Pays-Bas dans l'objectif de fournir un moyen technologique solide pour l'annotation et l'exploitation des enregistrements multimédias. Il est spécialement conçu pour l'analyse de la langue, la langue des signes et le geste, mais il peut être également exploité à d'autres fins, comme l'annotation et l'analyse des corpus des média.

[1] http://ircom.huma-num.fr/site/p.php?p=ressourceslogiciels
[2] Disponible sur :
http://icar.univ-lyon2.fr/ecole_thematique/idocora/documents/Abouda-Baude-ESLO.pdf

Elan est téléchargeable gratuitement à l'adresse suivantes : https://tla.mpi.nl/tools/tla-tools/elan/download/.

Quant aux instructions à suivre pour son installation et au guide complet de son utilisation en anglais, on peut les trouver dans le site : https://tla.mpi.nl/tools/tla-tools/elan/. Des guides en français peuvent être téléchargés aux sites suivants :
(http://llacan.vjf.cnrs.fr/res_manuels.php) et
(http://perso.ens-lyon.fr/isabelle.colondecarvajal/pro/wp-content/uploads/2015/02/ElanICC.pdf).

Elan est doté de plusieurs fonctionnalités, comme l'affichage de signaux audio et/ou vidéo avec leurs annotations, la création d'un nombre illimité de niveaux d'annotation que l'utilisateur peut définir selon son besoin, la fusion de différentes annotations d'un fichier média opérées dans des fichiers Elan séparés et entre autres beaucoup d'options de recherche.

II. Présentation de l'interface d'Elan

Chaque projet Elan est constitué d'au moins deux fichiers : un (ou plusieurs) fichier(s) média(s) (*.mpg, *.mov, *.wav, etc.) et un fichier d'annotation (*.eaf "EUDICO Annotation Format"). Le fichier d'annotation peut aussi être importé d'autres outils d'annotations, tels que Shoebox (*.txt ou *.cha) et Transcriber (*.SRT), et être sauvegardé sous forme d'un fichier Elan (*.eaf). Elan permet d'associer jusqu'à quatre fichiers vidéo à un seul et même document d'annotation.

Elan est un logiciel qui génère un fichier texte au format XML (*.eaf) renfermant les annotations crées par l'utilisateur, ainsi que les différentes données relatives aux fichiers audio/vidéo auxquels ces annotations sont associées. Voici comment son interface se présente :

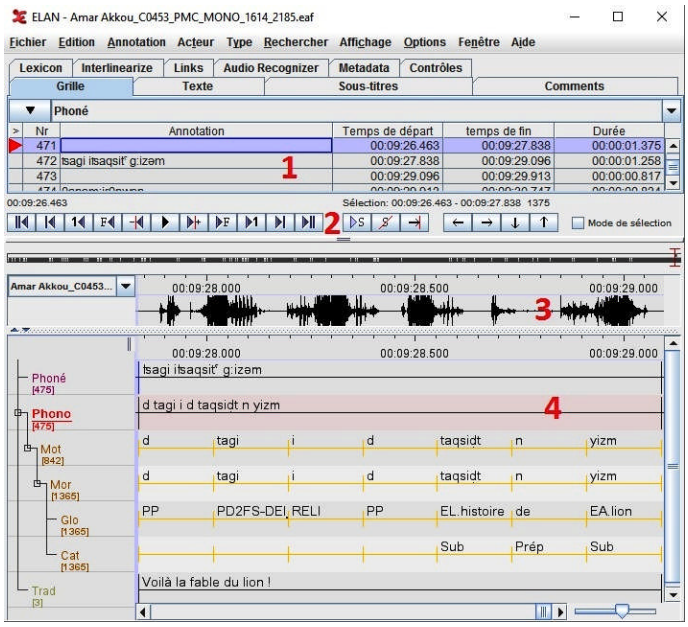

Figure 1 : La fenêtre d'Elan complète

1 : Zone de l'affichage des annotations sous différents aspects.
2 : Zone du contrôle et de la sélection audio ou vidéo.
3 : Zone de la wave form. Dans cette zone, il est possible de voir l'oscillogramme où on peut sélectionner le bloc temporel dont on annote le contenu.
4 : Zone de la partition de travail pour les annotations.

Une annotation peut être une notation d'un segment correspondant à des groupes prosodiques (groupes de souffle) ou sémantiques (phrases, syntagmes…), une glose, une traduction ou n'importe quelle description opérée par l'utilisateur ayant un rapport avec la source média. Les lignes d'annotations peuvent être engendrées sur des niveaux multiples, appelés *acteurs* (ou *tiers*).

Un acteur peut être hiérarchiquement dépendant (appelé "enfant de l'acteur"), c'est-à-dire qui contient des informations qui sont liées aux annotations d'un autre acteur (appelé "parent de l'acteur"), ou bien indépendant, qui contient des informations directement reliées à un intervalle de temps de paroles du locuteur.

III. L'annotation morphosyntaxique sous Elan

L'emploi d'Elan pour l'annotation morphosyntaxique de corpus oral synchronisé avec le son se fait par exemple, tel que le montre la figure 2 plus loin, en créant les lignes d'annotation suivantes :

- Une ligne **Phoné** pour noter le contenu oral de chaque segment en transcription phonétique, de type **"none"** (directement liée au temps), pas de parent ;
- Une ligne **Phono** pour noter le contenu oral de chaque segment en transcription phonologique, de type **"none"** (directement liée au temps), pas de parent ;
- Une ligne **mot** pour séparer chaque mot de la phrase, de type **"symbolic subdivision"** (soit ses sous-unités ne sont pas synchronisées temporellement), enfant de Phono ;
- Une ligne **Mor** pour segmenter les mots en morphèmes, de type **"symbolic subdivision"**, enfant de Mot ;
- Une ligne **Glo** afin de gloser chaque morphème, de type **"symbolic association"** (soit il y a une correspondance terme à terme entre l'annotation parent de l'acteur et celle qui s'y réfère dans l'enfant de l'acteur), enfant de Mor ;
- Une ligne **Cat** pour catégoriser grammaticalement chaque morphème, de type **"symbolic association"**, enfant de Mor ;
- Une ligne **Trad** pour la traduction libre du contenu oral du segment, de type **"none"**, pas de parent.

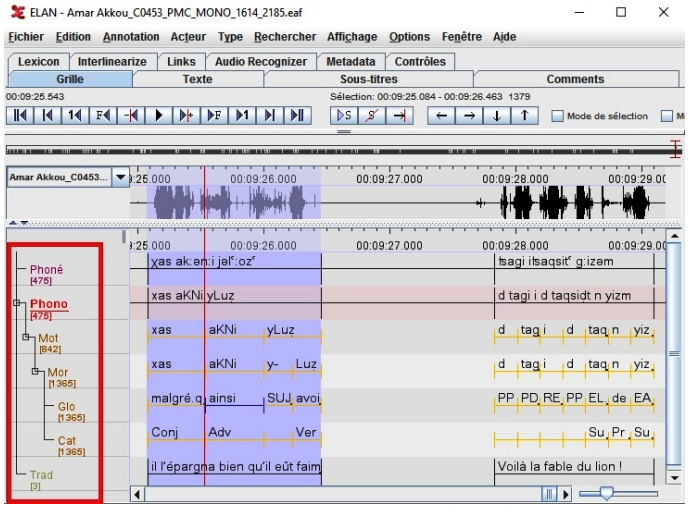

Figure 2 : Les acteurs (tiers)

Chaque mot d'une phrase se trouvant dans une ligne (mot) est alors segmenté en morphèmes dans une ligne en dessous (Mor), tout morphème étant ensuite glosé dans une 3ème ligne (Glo) et étiqueté grammaticalement dans une 4ème ligne (Cat). Durant cette opération d'interalignement, un alignement vertical est gardé entre d'une part chaque mot et le premier morphème dont il est composé et d'autre part entre chaque morphème, sa glose et sa catégorie grammaticale.

IV. Fonctionnalités supplémentaires d'ELAN-CorpA

ELAN-CorpA[1] est une version étendue d'Elan. Il a été développé par Coralie Villes et Christian Chanard du laboratoire du LLACAN[2], dans le cadre d'un projet financé par l'Agence Nationale de la Recherche (France) qui a pour objectif de mettre à disposition un corpus de langues afro-asiatiques (chamito-sémitiques) comportant une indexation texte-son et une annotation complexe. Ce projet est piloté par la linguiste berbérisante Amina Mettouchi.

A défaut d'absence d'un lexique d'appui, la segmentation et l'annotation d'un texte dans Elan ne peut se faire que manuellement ou bien en recourant à un autre logiciel, tel que Toolbox, qui permet d'effectuer un interalignement à partir d'un lexique et ensuite d'importer les données annotées vers Elan. L'idée d'ELAN-CorpA était donc de simplifier le processus d'interalignement dans Elan.

Afin de permettre la gestion d'un lexique au format XML et la segmentation interactive en morphèmes des mots d'une phrase, cette version d'Elan est doté d'un onglet supplémentaire appelé "Interlinearize". Cette nouvelle fonctionnalité permet d'assurer le découpage des unités en morphèmes et leur annotation s'effectue directement dans les lignes appropriées (Mor, Glo et Cat) du fichier Elan. Pour accélérer le processus d'annotation dans ELAN-CorpA, le lexique des mots segmentés et annotés dans des textes antérieurs peut être créé et importé afin d'enrichir celui d'un nouveau texte. La figure 3 ci-après montre cet onglet supplémentaire et le tableau du lexique sur lequel on peut s'appuyer afin d'optimiser le processus d'annotation.

[1] http://corpafroas.tge-adonis.fr
[2] http://llacan.vjf.cnrs.fr

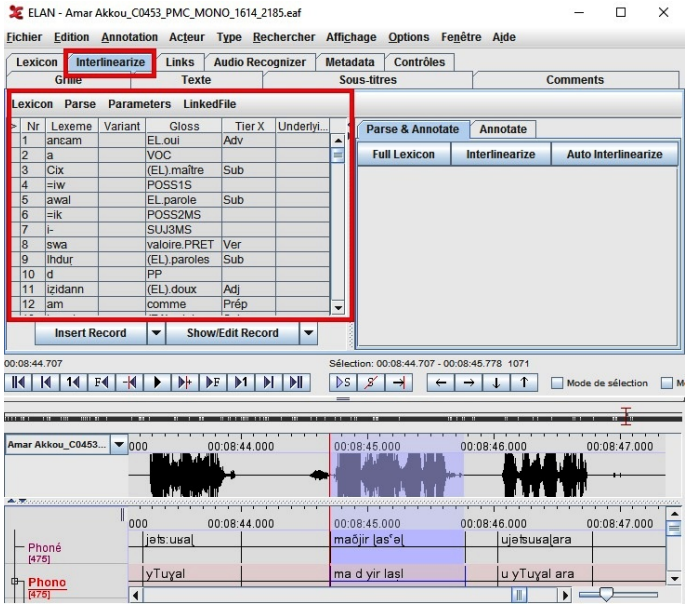

Figure 3 : Le lexique d'appui

Une entrée du lexique Elan peut être un lexème ou un affixe. Ce dernier est suivi ou précédé d'un tiret (-) ou du signe (=)[1] suivant qu'il s'agit respectivement d'un préfixe ou d'un suffixe.

V. Les options de recherche dans Elan

Les différentes données existant dans un/plusieurs fichier(s) annoté(s) dans Elan peuvent être recherchées au moyen d'un module qui offre la possibilité de trouver les occurrences de lexèmes ou de morphèmes dans les contextes spécifiés.

La liste des occurrences ainsi trouvée permet également de choisir un segment particulier pour l'écouter ou le voir entièrement. Par exemple, la figure 4 plus bas nous montre certaines occurrences qui concordent avec la glose « PROX » dans un extrait du corpus de notre thèse :

[1] Par exemple dans un corpus de berbère, on peut utiliser le tiret (-) pour les indices de personne verbaux et réserver le signe (=) pour les différents affixes de la langue.

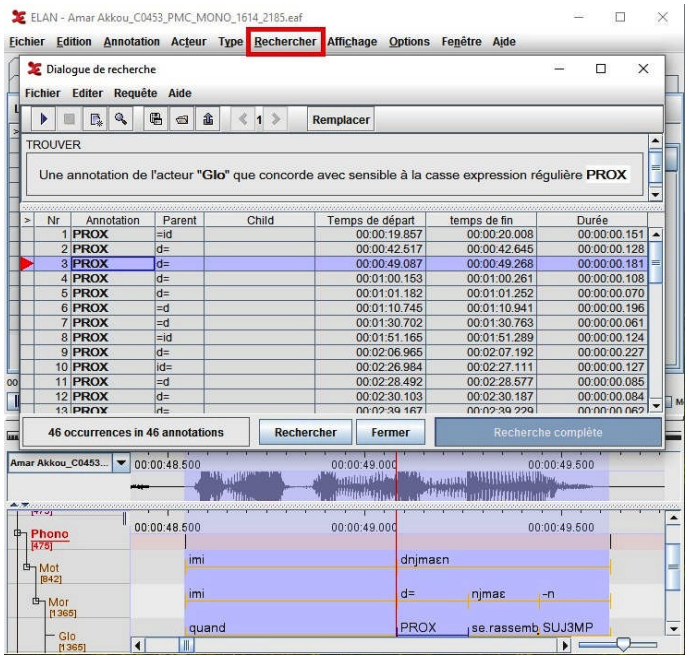

Figure 4 : Les options de recherche

Afin d'effectuer des recherches avec des conditions supplémentaires, il est toujours possible d'utiliser des expressions régulières. Une expression régulière est un modèle qui s'applique à un texte et permet de trouver toutes les portions du texte qui lui correspondent. La syntaxe des expressions régulières dans Elan peut être consultée au niveau de l'aide de la boite de dialogue "Recherche", comme le montre la figure 5 suivante :

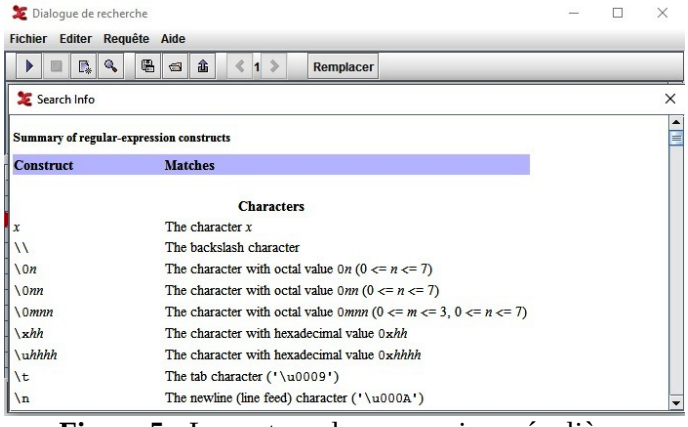

Figure 5 : La syntaxe des expressions régulières

VI. Les limites d'Elan

Elan nous donne la possibilité de réaliser des transcriptions directement sur l'audio en alignant des segments d'annotation avec des segments du signal sonore. Néanmoins, il n'offre pas une localisation très précise de la frontière du segment du signal sonore sur le fichier de transcription. Praat, contrairement à Elan, permet de corriger et d'approfondir des transcriptions en lien direct avec les données audio. Dans la mesure où les fichiers de transcription Praat (fichiers .TextGrid) peuvent être importés dans Elan, ils constituent donc un moyen pratique pour l'enrichissement des annotations dans Elan.

Bibliographie[1]

Abouda, Lotfi & Baude, Olivier, 2007 : « Constituer et exploiter un grand corpus oral, choix et enjeux théoriques le cas des Eslos »**,** in actes du colloque *Corpus en lettres et sciences sociales, des documents numériques à l'interprétation*, Colloque d'Albi Langages et Signification, juin 2006, Presses universitaires de Toulouse, pp. 161-168. Disponible sur : http://icar.univ-lyon2.fr/ecole _thematique/idocora/documents/Abouda-Baude-ESLO.pdf

Berkaï, Aziz, 2002 : *La terminologie linguistique en tamazight*, Magister de berbère, Université de Bejaia.

Chanard, Christian, 2014 : « Application du logiciel ELAN à l'annotation linguistique ». Disponible sur : http://llacan.vjf.cnrs.fr/fichiers/manuels/ELAN/ELAN_Annotation_linguistique.pdf

Colón de Carvajal, Isabel, 2013 : « Guide pratique pour annoter sous ELAN », Laboratoire ICAR, Université Lyon 2. Disponible sur : http://perso.ens-lyon.fr/isabelle.colondecarvajal/pro/wp-content/uploads/2015/02/ElanICC.pdf

Djemai, Salem, 2013 : *L'expression de la qualité en berbère, étude morphosémantique et syntaxique de l'adjectif en kabyle*, Thèse de Doctorat (dir. K. Naït-Zerrad), Paris, INALCO.

Habert, Benoît & alii, 1997 : *Les linguistiques de corpus*, Armand Colin, Paris.

Mettouchi, Amina, 2008 : « corpus oraux : des données à la théorie en passant par la technique », *Théories et données linguistiques*, Ecole d'été du CLI, 25-29 mai, Porquerolles, France.

Mettouchi, Amina & Chanard, Christian, 2010 : « From Fieldwork to Annotated Corpora : the CorpAfroAs Project », *Faits de Langue*, Les Cahiers n°2, pp. 255-265. Disponible sur : http://corpafroas.tge-adonis.fr/fichiers/Mettouchi_Chanard.pdf

Saad-Buzefran, Samia, 1996 : *Lexique d'informatique (français-anglais-berbère) Amawal n tsenselkimt*, l'Harmattan, Paris.

[1] Tous des liens internet indiqués dans cet article sont vérifiés le 22/12/2015.

Véronis, Jean, 2000 : « Annotation automatique de corpus : panorama et état de la technique », in Pierrel J-M (éd.), pp. 111-130.

Quelques liens utiles

http://corpafroas.tge-adonis.fr (ELAN-CorpA).

http://ircom.huma-num.fr (Le Consortium Corpus Oraux et Multimodaux).

http://llacan.vjf.cnrs.fr/res_manuels.php (Manuels Elan et autres).

https://tla.mpi.nl/tools/tla-tools/elan/ (Manuel complet d'Elan en anglais).

https://tla.mpi.nl/tools/tla-tools/elan/download/ (téléchargement d'Elan).