ELAN - CorpA

ELAN-CorpA est dérivé du logiciel ELAN utilisé pour annoter des enregistrements audio ou vidéo. Ici, l'annotation est assisté par un parseur et un lexique. Une ligne *ref* est créée pour référencer (numéroter) chaque segment. Chaque segment comportera ensuite successivement les lignes suivantes : une ligne *tx* pour la transciption, une ligne *mot* dans laquelle chaque mot est isolé dans une cellule, une ligne *mb* de segmentation des mots en morphèmes, une ligne *ge* de glose et une ligne *ps* de catégorie pour chaque morphème et enfin une ligne *ft* pour la traduction libre.

Pour assurer une homogénéité des configurations au sein de l'équipe, les fichiers s**enel.typ**, **mdf.typ** seront préalablement copiés dans le sous-dossier *Settings* du sous-dossier *Toolbox* du dossier *Senelangues* ainsi que le fichier **Senel1.etf** dans le sous-dossier ELAN.

SENELANGUES
- AUDIO
  - Fichiers.wav…
- ELAN
  - **Senel1.etf**
  - Fichiers ELAN…
- TOOLBOX
  - Dictionnaire
  - Fichiers annotés Toolbox…
  - SETTINGS
    - **Mdf.typ**
    - **senel.typ**

Création d'un fichier **ELAN**

ELAN génère un fichier XML contenant les informations de segmentation et d'annotation d'un fichier *média* (audio ou vidéo). Il faut donc commencer par charger un fichier *media*.

Créer un nouveau Document ELAN

- FICHIER > NOUVEAU
- *Fichier du type* : **Media files**
- *Rechercher dans :* Sélectionner le fichier audio (**.wav**)
- Cliquer sur le bouton >> entre les 2 fenêtres
- *Fichier du type* : **Template**
- *Rechercher dans :* Sélectionner le fichier modèle (**.etf**)
- Cliquer sur le bouton >> entre les 2 fenêtres
- OK

Enregistrer le fichier ELAN

- FICHIER > ENREGISTRER SOUS : nom_du_fichier, ENREGISTRER

  le nom du fichier sera du type **SEN_codelangue_InitialesChercheur_type_num**
  type = conv(ersation) ou narr(ation) ; num = numéro du fichier dans la série

Supprimer le rang *Default*

- CLIC-DROIT sur l'étiquette *Default*
- SUPPRIMER *Default*
- OK

Importer les rangs *ref, tx, mot* et *ft*

Si vous n'avez pas ouvert le modèle en même temps que le fichier audio :

- ACTEUR > IMPORTER ACTEURS
- CLIQUER sur le bouton RECHERCHER et sélectionner le fichier *Senel1.etf*
- CLIQUER sur SELECT, puis IMPORTER, puis FERMER

Ranger les rangs dans l'ordre voulu

- Cliquer sur l'étiquette du rang à déplacer et maintenir appuyé le bouton de la souris
- Déplacer l'étiquette à l'emplacement voulu, puis relâcher

Ranger les rangs dans l'ordre hiérarchique

- Cliquer avec le bouton de droite dans la zone des étiquettes des tiers
- SORT TIERS > SORT BY HIERARCHY

Segmentation du fichier *Audio*

Trois méthodes: 1) la segmentation au kilomètre qui permet en une fois de définir les segments voulus. La frontière de ces segments pourra être corrigée par la suite. 2) la segmentation progressive avec annotation au fur et à mesure. 3) segmentation automatique sur la base des silences (pause).

1 - Segmentation au kilomètre

- OPTIONS > SEGMENTATION MODE
- Vérifier que la ligne de segmentation sélectionnée est bien ***ref*** (en rouge)
- SELECTIONNER '*Une frappe par annotation*' (chaque fois qu'on tapera la touche <Entrée>, il y aura une frontière de segment)
- CLIQUER sur le bouton LIRE LA SELECTION 
- TAPER la touche <ENTREE> à la fin de chaque unité à segmenter
- CLIQUER sur le bouton PAUSE  pour arrêter la segmentation

Les segments sont délimités dans le rang ***ref***. La segmentation peut être interrompue et reprise ultérieurement par ce même processus.

2 - Segmentation au fur et à mesure

- CLIQUER au début de la zone à segmenter pour y positionner le fil rouge
- COCHER la case MODE DE SELECTION
- CLIQUER à la fin du segment voulu (la zone apparaît plus sombre) et l'écouter avec le bouton LIRE LA SELECTION 
  - Pour repositionner la frontière de droite, Cliquer à l'endroit voulu
- Une fois la zone à transcrire correctement définie, faire un CLIC-DROIT dans cette zone, **au niveau du rang *ref***, et sélectionner NOUVELLE ANNOTATION ICI. Une zone de texte s'ouvre. La fermer avec CTRL/ENTER. (le contenu du rang ref sera généré automatiquement, voir plus bas)
- Pour saisir la transcription, DOUBLE-CLIQUER sur le rang ***tx*** en dessous du segment ***ref*** correspondant. Fermer la fenêtre de saisie par CTRL/ENTER.
- Pour saisir la traduction libre, DOUBLE-CLIQUER sur le rang ***ft*** en dessous du segment ***ref*** correspondant. Fermer la fenêtre de saisie par CTRL/ENTER.

Délimiter un nouveau segment à la suite du précédent

- VERIFIER que la case MODE DE SELECTION est COCHEE
- CLIQUER sur le bouton  , le fil rouge se déplace en fin de segment
- CLIQUER sur le bouton  pour désélectionner le segment précédent
- CLIQUER à la fin du segment voulu (la zone apparaît plus sombre)
  - Pour repositionner la frontière de droite, Cliquer à l'endroit voulu
- CLIQUER sur le bouton  pour écouter le segment sélectionné
- Une fois le segment à transcrire correctement défini, faire un CLIC-DROIT dans cette zone, **au niveau du rang *ref***, et sélectionner NOUVELLE ANNOTATION ICI.

3 - Segmenter automatiquement à partir des silences (pauses)

Un outil de segmentation automatique basé sur des *échantillons de silence* est disponible. Il permet de créer une tier comportant des segments de signal et des segments de silence qui pourra servir de base pour créer la tier de segmentation du son en unités prosodiques (tier 'ref')

- Onglet AUDIO RECOGNIZER
- RECOGNIZER: Sélectionner : SILENCE RECOGNIZER MPI-PL
- Sélectionnez quelques millisecondes (~30 ms) de silence sur le signal audio
- Cliquer sur le bouton plus (+) dans la zone SELECTION PANEL, pour ajouter cet extrait dans la fenêtre Séléctions. Recommencer l'opération 3 ou 4 fois sur des zones de silences différentes
- Dans la zone SETTINGS : Choisissez la durée minimale d'un silence (200~250 ms par exemple) et la durée minimale d'un non-silence (par exemple 80~100 ms)
- Dans la zone PROGRESS, cliquez sur le bouton START

Des barres verticales apparaissent dans la zone du signal audio, délimitant les zones de signal marquées d'un 'x' et les zones de silences marquées d'un 's'

Si vous n'êtes pas satisfait de la segmentation, Ajoutez (+) ou enlevez (-) des échantillons de silence dans la zone *Selections* ou changez les valeurs minimales de silences et non-silence dans la zone *Settings*, puis relancez **Start**. Lorsque vous êtes satisfait,

- Cliquez sur CREATETIERS
- Sélectionnez l'onglet ALL SEGMENTATIONS
- Cliquez sur CREATE, puis CLOSE

Une tier *Channel1* contenant les segments de silence 's' et de signal 'x' a été créée. Cette tier est de type 'segmentation'. Pour vous servir de cette ligne comme ligne de base dans ELAN-CorpA, vous devez changer son type en '*ref* ', ainsi que son nom en '*ref* ' (cf. 17).

La durée d'un segment peut être lue en le sélectionnant et en lisant cette valeur au dessus du bouton de sélection représentant une flèche pointant vers le haut. (Cette valeur pourra être reportée comme annotation d'une Pause par exemple)

Déplacer la frontière entre deux segments

- Pour déplacer une frontière vers la droite, Sélectionner le segment de gauche en CLIQUANT à l'intérieur du segment, au niveau du rang *ref*. (Pour déplacer une frontière vers la gauche, Sélectionner le segment de droite.)

- Maintenir appuyée la touche ALT et amener le curseur sur la ligne de partage entre les deux segments, au niveau du rang *ref*.
- Cliquer (le curseur se transforme en une double-flèche) et déplacer la frontière vers la droite (ou la gauche), puis relâcher.

Repositionner les frontières des segments

- OPTIONS > SEGMENTATION MODE
- Vérifier que la ligne de segmentation sélectionnée est bien celle voulue (en rouge)
- Amener le curseur sur le segment à repositionner qui apparaît surligné en vert, le curseur en croix se transforme en une flèche lorsqu'on le situe vers une frontière, cliquer et déplacer la souris dans le sens voulu.

Attention : le déplacement d'une frontière peut déplacer les segments suivants ou précédents si le mode défini dans OPTIONS, PROPAGATION DU TEMPS est *bulldozer* ou *décalage*

Mode bulldozer : Tous les segments suivants ou précédents sont décalés de la valeur du déplacement de la frontière

Mode décalage : Tous les segments suivants ou précédents sont décalés, mais les espaces vides sont d'abord occupés, le décalage des segments suivants ou précedents étant réduit d'autant.

Créer un segment entre deux segments déjà existants

Pour ajouter un segment entre deux déjà existants

- Sélectionner l'espace correspondant au segment à rajouter, à cheval entre les deux segments existants
  - Pour repositionner la frontière de droite, Cliquer à l'endroit voulu *en maintenant la touche majuscule appuyée*
- Clic-droit sur l'espace sélectionné, au niveau de la tier ref et sélectionner NOUVELLE ANNOTATION ICI.

Annotation de segments prédéfinis

Si vous n'avez pas saisi les transcriptions au fur et à mesure

- DOUBLE-CLIQUER dans le segment à annoter, au niveau du rang ***tx*** ou ***ft*** . Une zone de texte s'ouvre
- TAPER le texte voulu (en vernaculaire dans *tx*, en français dans *ft*) et terminer par CTRL/ENTREE.

Pour remplacer la combinaison de touche CTRL/ENTREE par simplement ENTREE
- EDITION > PREFERENCES > EDITER PREFERENCES > EN TRAIN D'EDITER
- COCHER "la touche entrée valide les changements…"

Labeliser le rang *ref* (référencer les segments)

Il s'agit de numéroter automatiquement les segments, avec une étiquette (label)

- ACTEUR > LABELISER ET NUMEROTER…, Sélectionner **ref**
  - *Inclure la partie label* : **codelangue_InitialesChercheur_type_num**
    type = *conv*(ersation) ou *narr*(ation) *;* num = *numéro du fichier dans la série*
  - *Insérer autre délimiteur* : TAPER _ (caractère souligné)
  - OK puis FERMER

Tokeniser le rang *tx* dans le rang *mot* (une cellule par mot)

Il s'agit d'isoler chaque mot de la ligne *tx* dans une cellule individuelle

- ACTEUR > TOKENISER L'ACTEUR
    - *Source* : **tx**
    - *Destination* : **mot**
    - COMMENCER puis FERMER

Exportation vers Toobox

- FICHIER > EXPORTER VERS : **Fichier Shoebox**
    - *Utiliser le type de base de données Shoebox* : Cliquer sur [...], Rechercher **senel.typ** dans le sous-dossier *Settings* du dossier Toolbox
    - Cocher *Encoder les marqueurs en Unicode*, OK
- Rechercher le dossier destinataire, donner un nom au fichier Toolbox, ENREGISTRER

TOOLBOX

Ouvrir le fichier issu de l'exportation ELAN

- FICHIER > OUVRIR
- Rechercher le dossier voulu, Sélectionner le fichier issu de l'exportation de ELAN
- OUVRIR
- Lancer l'interalignement (voir document *TOOLBOXELAN.PDF*)

RETOUR DANS ELAN

Ouvrir le fichier annoté avec TOOLBOX dans ELAN

- Fermer le projet TOOLBOX
- Ouvrir ELAN
- FICHIER > IMPORTER > FICHIER SHOEBOX
- Sélectionner le fichier Toolbox en cliquant sur les 3 petits points [...]
- Cocher TOUS LES MARQUEURS SONT UNICODE
- OK
- A la demande, Rechercher le fichier Wav correspondant au texte annoté
- FICHIER > ENREGISTRER SOUS : nom_du_fichier > ENREGISTRER

Procédure d'interalignement dans ELAN

Le processus d'interalignement tel que nous l'entendons, est celui par lequel les mots d'une phrase contenu dans une ligne (*mot* par exemple) sont segmentés en morphèmes dans une ligne en dessous (disons *mb*), chaque morphème étant ensuite glosé dans une 3ème ligne (*ge*) et étiquetté grammaticalement dans une 4ème ligne (*rx*). Durant ce processus, un alignement vertical est maintenu entre d'une part chaque mot et le premier morphème qui le constitue et d'autre part entre chaque morphème, sa glose et sa catégorie. Jusqu'à présent, ELAN ne permettait pas de s'appuyer sur un lexique pour segmenter et annoter un texte directement. Les seules solutions possibles étaient :

- d'opérer manuellement au découpage des unités et de remplir chacune d'elle
- de transférer le texte à annoter d'ELAN vers Toolbox qui lui, permet d'effectuer un interalignement à partir d'un lexique, et enfin de réimporter les données annotées dans ELAN.

L'idée était donc de simplifier la démarche en permettant à l'utilisateur d'effectuer un processus d'interalignement directement dans ELAN c.à.d que le découpage des unités en morphèmes et leur annotation s'effectuent directement dans les lignes appropriées (*mb*, *ge* et *rx*) du fichier ELAN.

Pour ce faire, un nouvel onglet « *Interlinearize* » a été créé.

Après avoir ouvert un fichier à annoter

- *Cliquer sur l'onglet « Interlinearize ».*

Création, ouverture, importation d'un lexique

La procédure d'interalignement repose sur l'existence d'un lexique qui peut être

- un tout nouveau lexique ELAN
- un lexique ELAN existant
- un dictionnaire Toolbox importé dans ELAN

Un menu « *Lexicon* » permet de choisir

- *Cliquer sur l'onglet «Lexicon »*

Créer un lexique ELAN

En choisissant *Create*, une fenêtre de sélection de fichiers s'ouvre. Choisissez le dossier dans lequel vous voulez enregistrer votre lexique et donnez-lui un nom. L'extension **.eafl** lui sera automatiquement ajoutée.

A l'écran s'affiche alors sur la gauche, un tableau contenant les différentes colonnes du dictionnaire, et sur la droite des éléments servant à l'interalignement et à l'édition du lexique (« Insert Record », « Insert Variant »... cf. fig. 3)

Ouvrir un lexique ELAN

En choisissant *Open*, une fenêtre de sélection de fichiers s'ouvre. Choisissez le dossier dans lequel se trouve votre lexique (d'extension **.eafl)** sélectionnez-le, puis ouvrez-le.

Importer un lexique Toolbox

Une fenêtre s'ouvre permettant de choisir comme source de données un fichier dictionnaire Shoebox/Toolbox (.txt) ou un fichier dictionnaire ELAN (.eafl).

- *Choisir Fichiers du type « text files (*.txt) »*
- *Sélectionner un fichier dictionnaire issu de Shoebox/Toolbox*

- *Cliquer sur « Select »*

Parmi les champs contenus dans chaque fiche du dictionnaire Toolbox, certains n'entrent pas en jeu dans le processus d'interalignement (c'est le cas des exemples, définitions etc.). ELAN ne traite que les concepts suivants (colonne de droite):

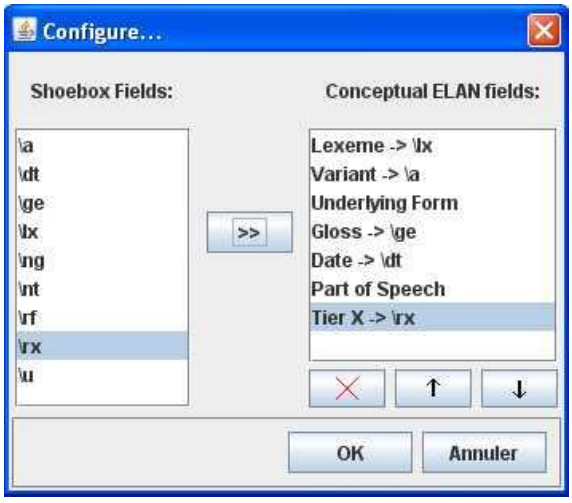
- **Lexeme** (toute entrée dans le lexique: lexème, affixe, mot-forme),
- **Variant** (forme alternative du lexème en contexte),
- **Underlying form** (décomposition de l'entrée en ses constituants, lorsqu'il y a assimilation par ex.),
- **Glose** (sens du lexème),
- **Part of speech (**catégorie grammaticale. Dans Toolbox, la catégorie (\ps) d'un mot recouvre les différents sens du mot. Ici, chaque sens (Gloss) d'un même mot du lexique aura sa propre catégorie (Tier X),
- **Tier X** (catégorie associée à l'entrée),
- **date** (de dernière modification de l'entrée).

Ces concepts doivent être liées aux champs présents dans le fichier Toolbox (colonne de gauche) afin d'être correctement traités.

Il est impératif que Lexeme, Glose et Tier X (ou Part of Speech) soient associés à un champ du fichier Shoebox/Toolbox. Toutefois, si vous n'utilisiez pas de champ \rx dans Toolbox, associez le champ de partie du discours de Toolobx (\ps) au champ Part of Speech de ELAN. L'importation copiera le contenu du champ Toolbox dans le champ ELAN Tier X.".

L'utilisateur choisit l'association voulue en appariant un à un les champs de droite aux champs de gauche.

- *Sélectionner un champ dans la liste de droite*
- *Sélectionner un champ issu de votre fichier Toolbox dans la liste de gauche*
- *Cliquer sur la flèche située entre les deux listes*



Le champ de droite sélectionné comporte maintenant une flèche '->' suivi du champ Shoebox/Toolbox approprié)

Pour supprimer une correspondance, sélectionner dans la colonne de droite l'élément voulu, puis

- *Cliquer sur la croix rouge*

Pour déplacer la correspondance sur l'élément situé au-dessus, puis

- *Cliquer sur la flèche vers le haut*

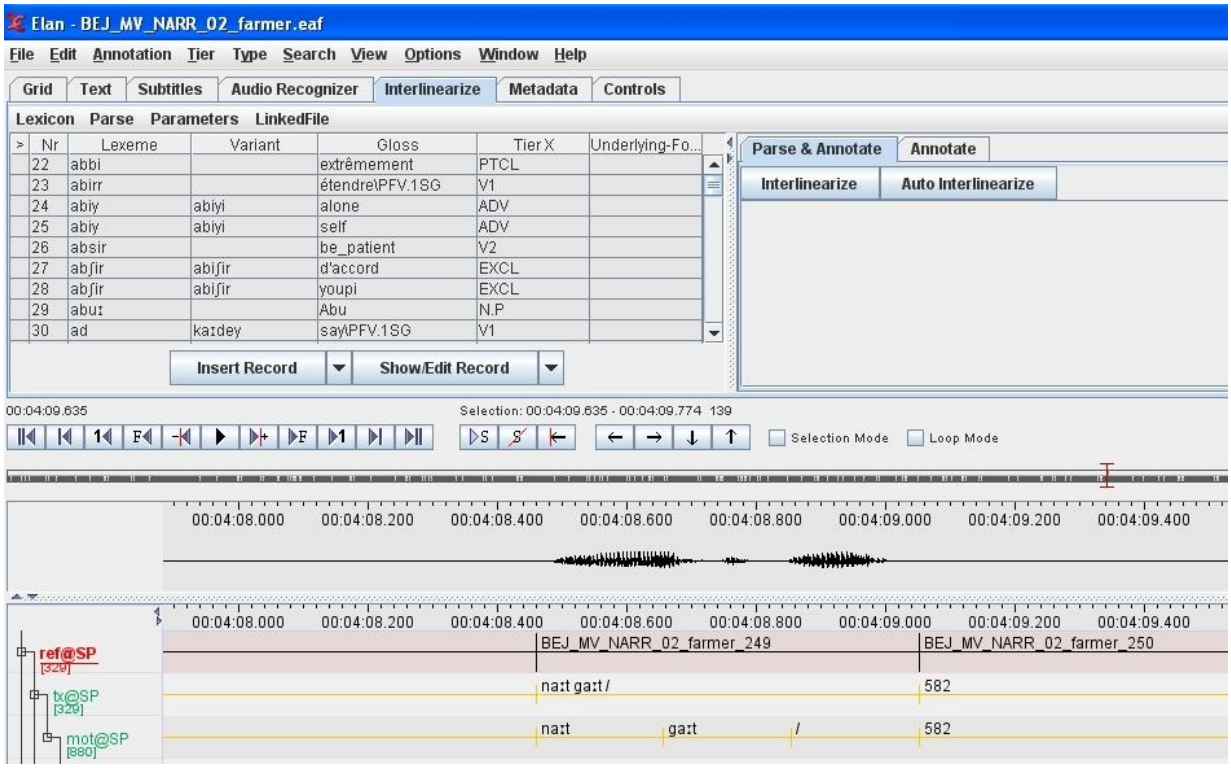
Pour déplacer la correspondance sur l'élément en dessous, puis

- *Cliquer sur la flèche vers le bas*

Lorsque toutes les correspondances nécessaires sont faites, cliquer sur « OK ».

Le programme récupère les éléments pertinents pour l'interalignement qu'il copie dans un nouveau fichier lexique ELAN (auquel il faudra donner un nom); c'est ce fichier XML nouvellement créé (d'extension **.eafl**) que le programme utilisera comme lexique pour l'interalignement (pour insérer de nouveaux morphèmes, modifier des éléments etc.).

A l'écran s'affiche alors sur la gauche, un tableau contenant les données du lexique, et sur la droite des éléments servant à l'interalignement et à l'édition du lexique (« Insert Record », « Insert Variant »...).



Ouverture automatique du lexique

Sur la droite des menus de la section *Lexique*, se trouve le menu *Linked File*. Celui-ci permet de choisir de rouvrir automatiquement ou non le lexique actuel au prochain démarrage de ELAN. Par défaut la case devant le nom du lexique ouvert est cochée, il se rechargera donc automatiquement..

Paramétrer l'interalignement

Avant de lancer l'annotation interactive des mots de la ligne voulue, il faut choisir cette ligne et définir les lignes d'annotation. Par défaut ces lignes sont « mot » pour la ligne contenant les mots à segmenter et annoter, « morph » pour la ligne de segmentation en morphèmes, « glose » pour la glose des morphèmes et « catmorph » pour l'étiquetage grammatical des morphèmes. Si ces lignes existent déjà, les annotations qu'elles contiennent déjà seront écrasées au fur et à mesure que le processus d'interalignement se fera. Ce paramétrage est à faire une fois pour toute.

Si ces lignes d'annotation n'existent pas encore dans votre fichier ELAN, faites:

- *Parameters, Interlinearize tier Parameters, configure interlinear Tiers*
- Choisissez la ligne source à segmenter et annoter : *Choose interlinear tier* (par ex *mot*)
- Cliquez sur *Ok*
- Choisissez les étiquettes de la ligne des segments (*morph*), de la glose (*glose*) et de la catégorie (*catmorph*)
- Cliquez sur le bouton *Create tiers*

Les lignes sont créées, le processus peut commencer.

Si ces lignes existent déjà dans votre fichier ELAN, faites:

- *Parameters, Interlinearize tier Parameters, Rename Interlinear Tiers*
- Donnez le nom de vos tiers
- Cliquez sur le bouton *Rename tiers*

Principes de l'annotation

Une entrée du lexique ELAN (que nous appelons un *Lexème)*, peut être un *Lemme* (forme de base choisie pour représenter l'ensemble des formes du mot en contexte) – présentant éventuellement des formes alternatives appelées ici v*ariantes* – ou un *Affixe*. Les affixes présentent un tiret (-) à gauche ou à droite suivant qu'il s'agit respectivement d'un suffixe ou d'un préfixe.

Le parseur commence par rechercher le mot en cours dans le lexique, au niveau du champ *Lexeme* ou du champ *Variant*. S'il est trouvé dans le champ *Lexeme*, il est présenté tel quel dans la section *Segmentation*. Sil est trouvé dans un champ *Variante*, c'est la forme lemmatisée du champ *Lexeme* qui est présentée dans la section *Segmentation*. L'entrée (ou les entrées en cas de polysémie ou d'homonymie) trouvée dans le lexique s'affiche dans la section *Lexique*. Un double-clic sur l'entrée du lexique valide l'annotation qui est reportée dans les lignes correspondantes du texte, sous le mot en cours. Le mot suivant du texte est sélectionné et le processus peut recommencer.

Si le parseur ne trouve pas le mot dans le lexique, il va commencer à rechercher une correspondance entre la fin ou le début du mot et les différents affixes du lexique. Chaque fois qu'il trouve une correspondance, il isole l'affixe trouvé et relance la recherche sur le segment restant (qui est alors traité comme un mot). Toutes les segmentations possibles s'affichent, en fin d'analyse, dans la section *Segmentations*. Le choix de l'une d'elles (par un double-clic dessus) réduit l'affichage du lexique aux seuls morphèmes de la segmentation choisie. Le choix de l'entrée lexicale correspondant à un morphème se fait en double-cliquant sur l'entrée du dictionnaire souhaitée (en cas de polysémie ou d'homonymie), le segment suivant étant alors automatiquement sélectionné dans la section *Segmentation*. Le dernier segment validé provoque le transfert des annotations dans le texte, et le mot suivant du texte est sélectionné et le processus peut recommencer.

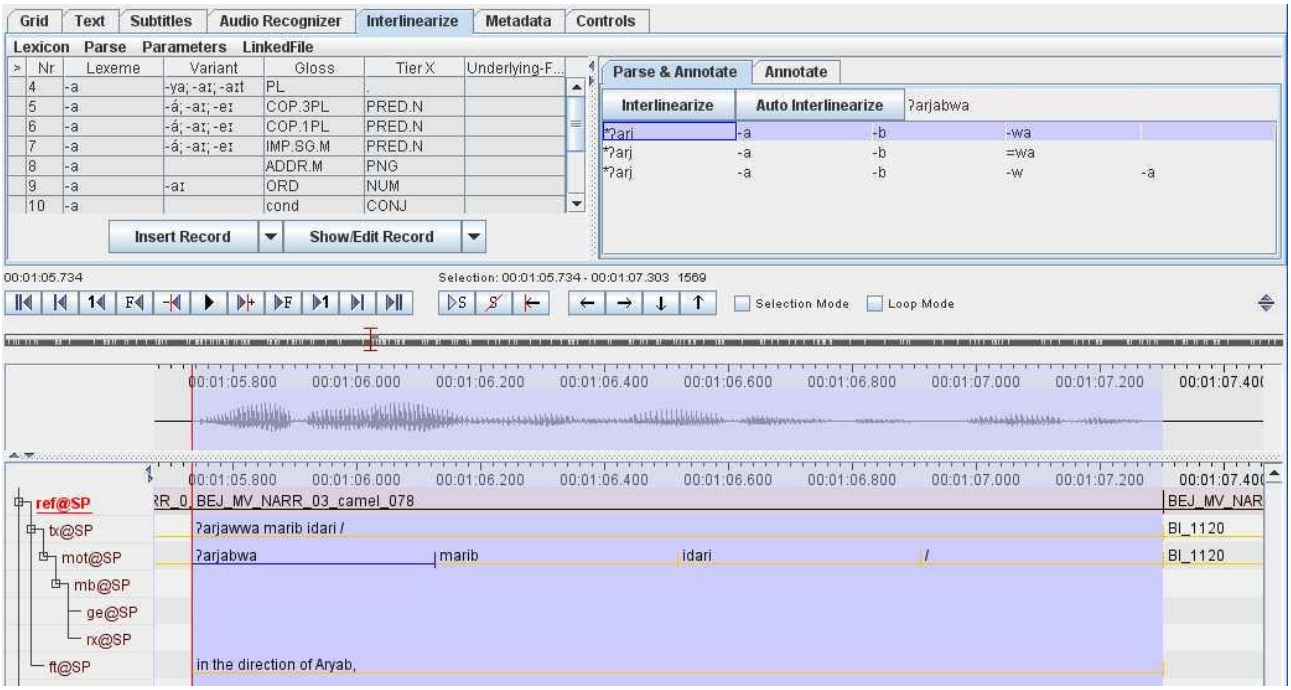
Lorsqu'un mot (ou le reste d'un mot, une fois les affixes connus isolés) n'est pas trouvé dans le lexique et que le parseur ne peut pas le segmenter sur la base des affixes du lexique, il apparaît dans la section *Segmentations* précédé d'un astérique (*). A ce stade, si ce mot comporte encore des affixes, il faut commencer par eux. Ensuite seulement le radical sera entré, soit tel quel, soit sous forme de variante d'un lemme.

Lancer le processus d'interalignement

Le parseur va chercher chaque mot à annoter dans le lexique, et s'il ne le trouve pas, il va essayer de lui trouver des segmentations possibles en fonctions des affixes contenus dans le lexique.

- Placer le curseur sur le premier mot de la ligne à segmenter, il se souligne en bleu.
- Cliquer sur le bouton « *Interlinearize* » dans la section ***Segmentations***, à droite du lexique.

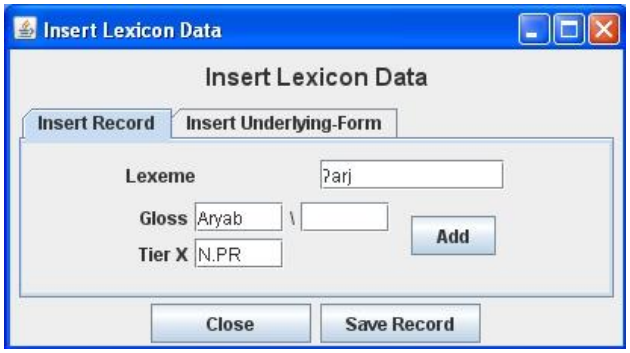
Les différentes segmentations possibles s'affichent dans la section *Segmentations*, et le lexique se réduit aux seuls segments intervenant dans les segmentations trouvées. *Le dernier segment non segmentable est précédé d'un **astérique** s'il n'existe pas dans le lexique*.



Dans l'exemple ci-dessus, le mot **ʔarjabwa** presente 3 segmentations possibles. Les suffixes **-a**, **-b** et **-wa** trouvés dans le lexique conduise à l'isolation d'un radical possible ***ʔarj**.

Saisie d'une entrée et/ou d'une variante dans le lexique (*Insert record*)

Pour entrer un mot nouveau dans le lexique, qu'il s'agisse d'un lemme ou d'un affixe, on peut cliquer dans section *Segmentations* sur le bouton « *Insert record* », ou bien



- Clic-droit sur le mot précédé d'un astérisque (ici *ʔarj*)
- Clic sur *Insert record*

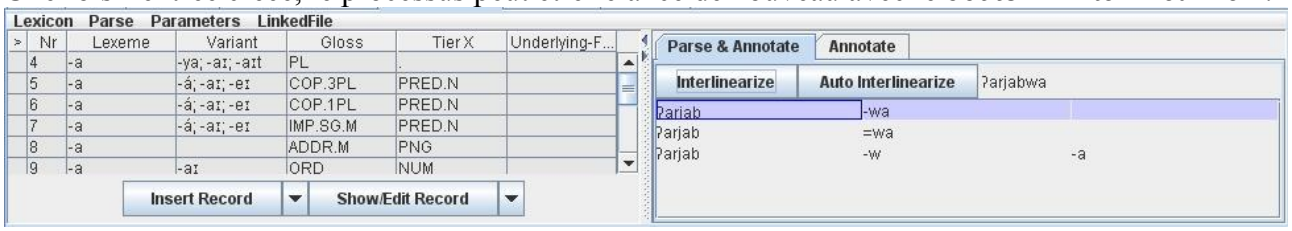
une fenêtre avec le mot choisi apparaît. Il peut être modifié. Par exemple, ici le mot à ajouter est *ʔarjab* glosé comme nom propre *Aryab*.

Sur la ligne de *Gloss*, on peut ajouter à droite du '\', une (des) étiquettes morphologiques correspondant à des valeurs de morphèmes non segmentables. (Par ex: \PL pour une forme plurielle d'un nom)

- Cliquer sur le bouton *Save Record*

Si le morphème que vous êtes en train de gloser contient d'autres éléments morphologiques qui ne peuvent être segmentés, ou que vous ne voulez pas isoler comme morphèmes séparés, vous pouvez utiliser la cellule à droite du symbole '\' pour entrer la valeur grammaticale de ces éléments. Remarquez qu'il ne faut pas entrer le délimiteur (\) devant l'étiquette grammaticale, celle-ci sera automatiquement ajoutée dans la ligne d'annotation.

Une fois l'entrée créée, le processus peut être relancé de nouveau avec le bouton « *Interlinearize* ».



Ici, avec le mot *ʔarjab* entré dans le lexique, 3 segmentations complètes sont possibles.

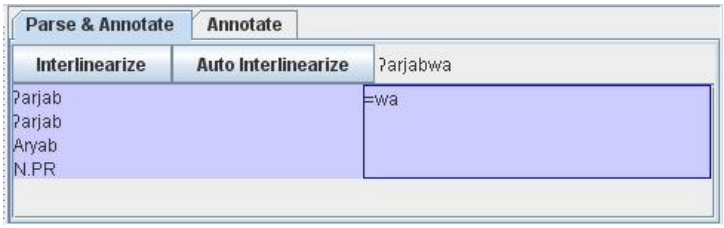
Sélection de la segmentation et de la glose

Il reste à faire le choix de la bonne segmentation

- Double-clic sur le premier segment de la ligne de segmentation voulue

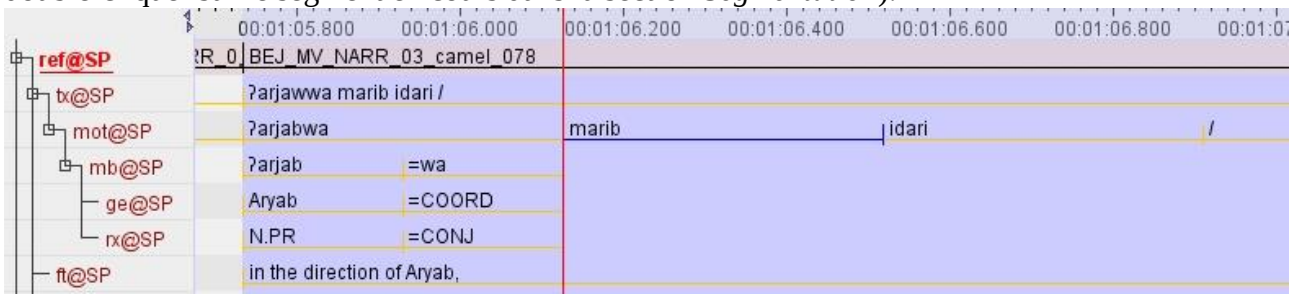
Le lexique se réduit à l'entrée correspondante (elles pourraient être multiples en cas de polysémie ou d'homonymie).

- Double-clic sur l'entrée du lexique voulue



L'annotation du premier segment s'affiche dans la section *Segmentation* et le segment suivant est sélectionné.

Et on recommence : un double-clic sur le segment sélectionné permet de réduire le lexique aux entrées correspondantes, puis un double-clic sur l'entrée voulue dans le lexique valide le choix et le segment suivant est sélectionné (pour gagner du temps, on peut aussi directement double-cliquer sur l'entrée du lexique voulue, sans double-cliquer sur le segment en cours dans la section *Segmentation*).



Lorsque le dernier segment a été annoté, les annotations choisies sont copiées dans leurs lignes correspondantes et le mot suivant est sélectionné

Affixes

Un affixe dans le lexique ELAN est une entrée qui comporte un tiret. Pour un préfixe on ajoutera un tiret (-) à la fin du segment et pour un suffixe au début.

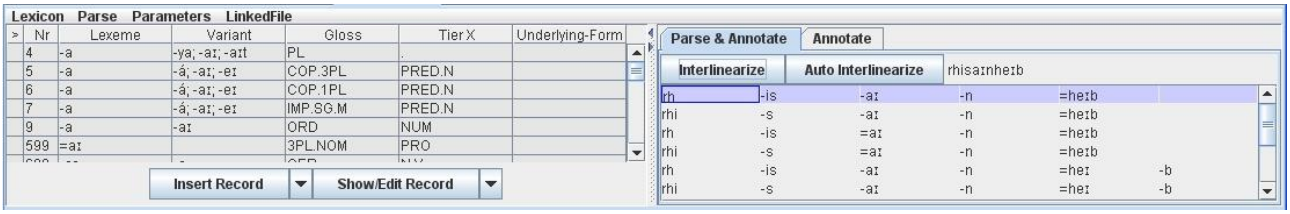
Pour entrer un affixe dans le lexique, on peut aussi faire un clic-droit sur le mot contenant cet affixe et choisir « *Insert a record* ». Dans la case de saisie de l'enregistrement, on supprimera ensuite tout sauf cet affixe.

En relançant le processus d'interalignement, une segmentation sera proposée isolant l'affixe du reste du mot. Ce reste est lui-même recherché dans le lexique. On recommencera autant de fois qu'il y a d'affixes dans le mot. Finalement le radical sera entré dans le lexique sous la forme d'un lemme ou d'une variante d'un lemme.

xtended features of the parsing

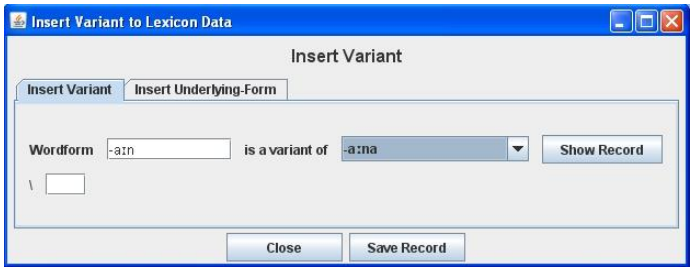
Morphophonolologie (lemme et variante)

Lorsqu'une transformation morphophonologique apparaît à la frontière d'un radical et d'un affixe (ou de deux affixes consécutifs), il faut toujours avoir à l'esprit que le parseur recherche une correspondance entre ce qui reste à segmenter et les entrées du lexique (au niveau des champs *Lexeme* et *Variant*)



Dans l'exemple ci-dessus, le parseur ne peut pas trouver la segmentation correcte du mot 'rhisaɪnheɪb' (rh –is -aɪna =heɪb) à cause de la disparition de la voyelle 'a' du suffixe '-aɪna' devant le clitique '=heɪb'. Lorsque '=heɪb' est isolé, pour que le parseur puisse correctement isoler le suffixe '-aɪna' qui se trouve déjà dans le lexique, nous pouvons entrer '-aɪn' comme variante de '-aɪna'.

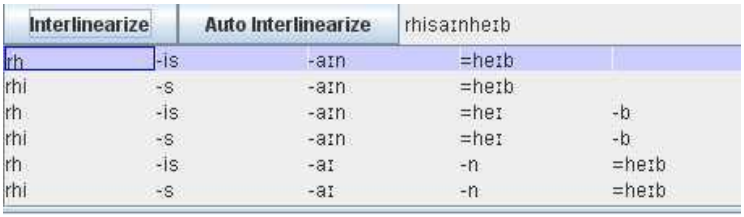
Ajouter une variante à une entrée



Remplacer le bouton « *Insert Record* » par « *Insert Variant* » grâce à la flèche déroulante, puis cliquer dessus.

- Entrer la forme variante (-aɪn) et sélectionner l'entrée correspondante (-aɪna).
- Sauvegarder l'enregistrement

Come le parseur recherche une correspondance (avec la fin ou le début du segment restant) au niveau des *lexemes* ou des *variantes* des lexèmes, il proposera maintenant la variante '-aɪn' de l'entrée '-aɪna' comme satisfaisant la correspondance.

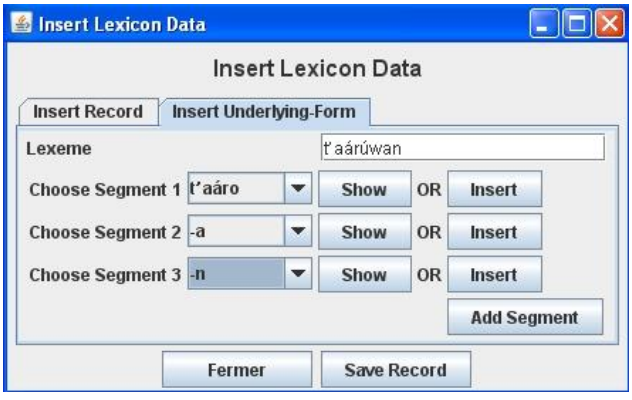


L'annotation peut maintenant continuer par la validation dans le lexique des différents segments du découpage; la validation du morphème '-aɪn' renverra la forme de base du *lexeme* '-aɪna' dans la ligne d'annotation *mb*.

Lorsque la morphophonologie est trop complexe pour permettre la segmentation correcte, même en ayant recours aux variantes des affixes ou radicaux, il est toujours possible de donner le découpage directement dans l'entrée lexicale.

Forme sous-jacente

Il peut arriver que la morphophonologie soit trop complexe et que la segmentation d'un mot soit difficile à gérer par l'ajoût de forme variante à une entrée lexicale. Dans ce cas, on peut donner directement le découpage du mot lors de la saisie de l'entrée lexicale. Mais attention, les différents segments composant le mot doivent alors exister dans le lexique.



- Clic sur le bouton *Insert Record* ou clic-droit sur le mot
- Sélectionnez l'onglet *Insert underlying form*
- Recherchez le premier segment (ici *t'aáro*)) en déroulant la liste en face de *Choose Segment 1*
- idem pour le segment 2, (ici -*a*).

Il est possible d'ajouter d'autre segment en cliquant sur le bouton *Add Segment*

- validez par *Save Record*, puis *Fermer*

En cas d'homonymie ou de polysémie dans les entrées lexicales, il peut être difficile de choisir, parmi plusieurs, le morphème voulu pour un segment donné. Le bouton *Show* permet d'afficher le contenu de l'entrée lexicale choisie pour vérifier qu'il s'agit bien de celle voulue.

De même, il est possible dans cette fenêtre d'ajouter une entrée qui ne serait pas déjà dans le lexique et dont on aurait besoin pour la segmentation du mot en cours.

- Cliquez sur le bouton *Insert* en face du segment en cours

une petite fenêtre *Insert Morpheme* s'ouvre permettant d'ajouter une entrée dans le lexique en la validant avec OK. Cette entrée s'affichera dnas le segment en cours.

Il faut bien remarquer que cette méthode qui consiste à fournir au parseur la segmentation had hoc d'un mot ne devrait pas être utilisée systématiquement, mais au contraire dans les seuls cas où le parseur n'arriverait pas à fournir la bonne segmentation sur la base des lexèmes (lemmes, variantes et affixes) que le lexique contient. En effet ce type d'entrée spécifique dans le lexique ne permet de segmenter que ce mot là, alors que le principe de base du parseur consistant à fournir d'une part des lemmes (avec d'éventuelles formes variantes) et d'autre part des affixes, est bien plus productif et systématique.

La fonction Auto Interlinearize

Afin de gagner du temps dans le processus d'interalignement, on peut choisir le traitement automatique d'un mot avec passage au mot suivant, lorsqu'une seule segmentation existe pour le mot en cours, sans ambiguïté sur les différents segments (une seule glose). Ainsi toute une séquence peut être traitée d'un coup jusqu'au prochain mot où un choix doit être fait par l'utilisateur.

Lancement de la fonction d'auto-interalignement

On peut lancer la fonction d'auto-interalignement à partir de n'importe quel mot de la ligne de base (dans notre exemple *mot*)

- Cliquer sur le premier mot où doit démarrer l'interalignement automatique (sa ligne de base devient bleue)
- Cliquer sur le bouton *Auto Interlinearize*

La segmentation et l'annotation du mot se fait, si les données du lexique le permettent, jusqu'au prochain mot où une ambiguïté ou un segment inconnu arrête le processus.

Lexique des segmentations

Un lexique des mots-formes et de leurs segmentations glosées (que nous avons appelé *Parse*) peut être exporté à partir des mot-formes trouvés dans un texte interaligné. Ce lexique peut être augmenté au fur et à mesure des nouveaux textes interalignés. Il pourra servir pour l'interalignement automatique d'un nouveau texte sur la base des segmentations et annotations déjà rencontrés dans de précédents textes interalignés. (On peut envisager également un éditeur de fichier *Parse* qui aiderait à traduire les gloses d'un lexique *Parse* dans une autre langue. On se servirait ensuite de ce nouveau lexique pour relancer l'interalignement automatique sur le même texte, ce qui produira en un temps record un texte annoté dans l'autre langue.)

Création, fusion, ouverture d'un fichier de segmentations

Pour exporter le lexique des mots-formes et de leurs segmentations, aller dans la section *Lexique* et choisir le menu *Parse* :

- *Parse, Export Parse data*
- Choisir le dossier destination et le nom du fichier. Une extension **.eafp** sera automatiquement ajoutée.

Pour fusionner le lexique des mots et de leurs segmentations avec un lexique *Parse* existant, aller dans la section *Lexique* et choisir le menu *Parse* :

- *Parse, Export Parse data*
- Choisir le dossier destination et le nom du fichier.

Pour ouvrir un fichier *Parse* existant afin d'utiliser la fonction d'auto-interalignement

- *Parse, Open Parse data*

Pour utiliser le lexique *Parse* dans le processus d'interalignement automatique

- Cocher la case *Search in Parse data file*
- Cliquer sur le bouton *Auto Interlinearize*

Ouverture automatique d'un lexique de segmentations

Par défaut, l'ouverture d'un fichier de segmentations (*Parse*) dans ELAN est mémorisée afin qu'une prochaine ouverture le charge automatiquement. Pour supprimer le lien entre le fichier ELAN en cours et un fichier de segmentations,

- Aller dans le menu *Linked File* dans la section *Lexique*
- Décocher la case devant le nom du fichier *Parse*

Le fichier *Parse* ne sera plus chargé à la prochaine ouverture du fichier ELAN.

Sauvegarde des fichiers liés

A la fermeture d'un fichier ELAN, une fenêtre s'ouvrira si un fichier lié n'a pas été préalablement enregistré. Par défaut, les fichiers liés seront enregistrés en quittant ELAN, mais en décochant la case devant le nom d'un fichier lié, on peut quitter ELAN sans le mettre à jour.