SHOEBOX version 4 - Seconde partie

Interalignement de texte

Définition

L'*interalignement de texte* est un processus plus ou moins automatisé consistant à placer de façon interactive des informations issues d'un ou plusieurs lexiques, sous chaque mot d'un texte. L'annotation du texte comprend généralement un découpage des mots en morphèmes (*segmentation*), une *glose* pour chaque morphème, et parfois la *catégorie grammaticale* à laquelle ils appartiennent. Une ligne de *traduction libre* peut être ajoutée à la main. L'interalignement de texte est une manière efficace d'enrichir un lexique tout en étudiant le morphologie de la langue. Voici un exemple de texte interaligné par Shoebox :

```
\id Tutoriel
\ref 1
\t Berne      en   opgroeid         yn   Ynje,
\m bern -e    en   op-  groei -e    yn   Ynje
\g bear -PSTP and  up-  grow  -PSTP in   Indonesia
\p V    -Tns  Conj Dir- V     -Tns  Prep N
\f Born and raised in Indonesia,

\ref 2
\t sil dêr syn grêf wêze.
```

Shoebox sait découper les mots en isolant les affixes suivant certaines règles plus ou moins simples, et générer le texte interaligné en recherchant les mots, les racines et les affixes dans un (ou plusieurs) lexique(s) et en renvoyant les informations contenues dans les champs appropriés définis par le paramétrage.

Importer du texte dans Shoebox

Shoebox peut importer du texte saisi au kilomètre, et créer soit une fiche par phrase, soit une seule fiche contenant tout le texte découpé en phrases. Dans les deux cas, un champ de référence est ajouté, contenant le numéro d'ordre de la phrase dans le texte d'origine.

Quelques préliminaires dans le traitement de texte d'origine

- Rajoutez en début de texte le marqueur **\name**, suivi d'un espace et d'un titre qui servira à repérer la provenance des phrases (par exemple le titre d'un conte).
- Enregistrer le fichier en format *Texte Seulement* sous un nom avec l'extension **.db** (par convention). (Dans Word, veillez à mettre ce nom entre guillemets sinon une deuxième extension **.txt** lui sera rajouté.)

Pour faire une fiche par texte

- Créez un *Type de base de données* avec comme marqueur d'enregistrement **\name**, s'il n'existe pas déjà
    - *Projet, Type de base de données, Ajouter...*
- Faites *Fichier, Ouvrir* et recherchez le fichier à importer

Ce fichier n'étant pas encore *estampillé* Shoebox, une fenêtre vous demande de lui attacher un type de base de données.

- Choisissez le type avec **\name** comme marqueur d'enregistrement
- Cochez la case *Utiliser Table CCT*
- Parcourez le dossier de configuration de Shoebox (MyShSet...) à la recherche du fichier **Textprep.cct** et sélectionnez-le
- Cliquez sur *OK*

Un message vous précisera que le fichier d'origine sera conservé sous le nom initial, avec l'extension **.ori** (pour original)

- Cliquez sur OK

Une fiche est créée avec un champ **\name** contenant le nom que vous avez donné au texte, un champ \t pour chaque phrase du texte, précédé d'un champ \ref contenant le nom du texte + le numéro de la phrase. Le découpage en phrases a été fait par la recherche du point, du point d'exclamation ou du point d'interrogation. (On peut rajouter le point-virgule ou tout autre marque en éditant le fichier *Textprep.cct*).

Pour faire une fiche par phrase

- Créez un *Type de base de données* avec comme marqueur d'enregistrement **\ref**, s'il n'existe pas déjà
    - *Projet, Type de base de données, Ajouter...*
- Faites *Fichier, Ouvrir* et recherchez le fichier à importer

Ce fichier n'étant pas encore *estampillé* Shoebox, une fenêtre vous demande de lui attacher un type de base de données.

- Choisissez le type avec **\ref** comme marqueur d'enregistrement
- Cochez la case *Utiliser Table CCT*
- Parcourez le répertoire de configuration de Shoebox à la recherche du fichier **Textprep.cct** et sélectionnez-le
- Cliquez sur *OK*

Un message vous précisera que le fichier d'origine sera conservé sous le nom initial, avec l'extension **.ori** (pour original)

- Cliquez sur OK

Une fiche est créée par phrase avec un champ **\ref** contenant le nom que vous avez donné au texte + le numéro de la phrase et un champ \t contenant la phrase. Le découpage en phrases a été fait par la

recherche du point, du point d'exclamation ou du point d'interrogation. (On peut rajouter le point-virgule ou tout autre marque en éditant le fichier *Textprep.cct*).

Numérotation

L'importation de texte à travers la moulinette *TextPrep* découpe et numérote les phrases du texte en vue de l'interalignement, avec des marqueurs de texte, de référence et de traduction libre standards (\t, \ref, \f). Le texte devra donc être importé suivant un type de base de données *texte* ayant le marqueur d'enregistrement \t.

Exemple de texte découpé par *TextPrep* :

Fryslân yn maitiidpracht, wylst de sinne skynde oer
de marren en de wide greiden mei fee. Noarwegen, doe't de hege sinne dreamde yn 'e fjorden.

\ref 1
\t Fryslân yn maitiidpracht, wylst de sinne skynde oer
de marren en de wide greiden mei fee.
\f
\ref 2
\t Noarwegen, doe't de hege sinne dreamde yn 'e fjorden.
\f

On vérifiera que le type de base de données *texte* servant à importer le document est bien configuré pour reconnaître ces champs pour d'éventuels redécoupage et renumérotation.

- *Base de données, Propriétés,* onglet *Numérotation*
- *Marqueur de référence* : Sélectionnez dans la liste déroulante le marqueur devant servir pour la référence
- *Marqueur de texte* : Sélectionnez le marqueur de la ligne de texte à découper.
- Cliquez sur *OK*

Pour pouvoir modifier ces paramètres, il faudra d'abord désactiver l'option de numérotation

- *Bases de données, Propriétés*, onglet *Numérotation*
- Cochez la case *Désactiver la numérotation des textes*
- Cliquez sur *OK*

Puis réactiver cette option

- *Bases de données, Propriétés*, onglet *Numérotation*
- Décochez la case *Désactiver la numérotation des textes*
- Choisir les champs voulus
- Cliquez sur *OK*

Remarque : Si l'interalignement a été paramétré pour ce type de base de données *texte*, on ne pourra pas modifier les marqueurs de texte et de référence servant à la numérotation.

Découper et numéroter du texte

Un texte **importé** dans Shoebox, a généralement été découpé par la table *TextPrep.cct* (voir plus haut). Mais on peut aussi saisir directement du texte dans Shoebox, ou insérer dans une fiche du texte mis dans le presse-papier par un *Couper-Coller*. Ce texte kilométrique pourra alors être découpé en phrases suivant des caractères de ponctuation définissables, chaque phrase se voyant précédée d'un champ de référence qui comportera un titre (également définissable) suivi du numéro d'ordre de la phrase dans le texte.

- Se positionner dans la fiche contenant le texte à découper
- *Outils, Découper/Numéroter du texte*
- Cochez *Enregistrement en cours* ou *Base entière* suivant le cas
- *Découper selon les ponctuations* : vérifiez les caractères de ponctuation servant de base pour la découpe en phrases ou locutions (rajouter , ou ; si nécessaire)
- Vérifiez le champ contenant le texte à découper. Eventuellement vous pouvez découper un autre champ que celui de préselectionné. (Pour cela il faut qu'il soit du même encodage de langue)
    - Cliquez sur *Choisir les champs*
    - Sélectionnez un à un les champs à découper et Cliquez sur *Ajouter*
- Comme titre de référence pour les phrases, vous pouvez choisir dans le cadre *Nom du texte* entre
    - *utiliser le contenu du champ* que vous sélectionnerez (name ou id...), ou
    - *utiliser le nom* que vous taperez dans la case en face
- Le numéro de départ de la numérotation peut être choisi en tapant ce chiffre dans la case *à partir du numéro*.

Cette commande pourra également servir avec du texte déjà découpé pour changer le titre de la réference ou pour produire un découpage sur une autre ligne.

Remarque : Si vous utilisez cet outil sur du texte précédemment importé par *TextPrep*, le champ de découpage par défaut sera le champ \t et le champ de référence par défaut sera \ref.

Renuméroter du texte

Si après découpage et numérotation, vous supprimez une phrase du texte, ou au contraire en rajoutez une, vous pouvez faire mettre à jour le champ référence

- *Outils, Renuméroter du texte*

Le projet

L'interalignement est un processus qui met en relation un texte à interaligner et un ou plusieurs lexiques dans lesquels les informations d'annotation seront recherchées. Il convient donc d'ouvrir dans le Projet en cours un ou plusieurs lexiques, en plus de la base de données de textes à interaligner. A titre d'exemple, soit le projet *frison* FRI2.PRJ (dans *Samples*) comportant :

- une base de données lexicales "FriRT.dic" de type *FrisanD* (marqueur d'enregistrement \fri)

```
\fri bern
\ps  V
\g   bear

\fri bikwaam
\ps  Adj
\g   capable
.....................
```

- une base de données *texte* "FriSampl.txt" de type *FrisanT* (marqueur d'enregistrement \id)

```
\id  Tutoriel
\ref 1
\t   Berne en opgroeid yn Ynje,
\f

\ref 2
\t   sil dêr syn grêf wêze.
\f
```

Visualiser les deux bases en même temps

- *Fenêtre, Mosaïque verticale*

Paramétrer l'Interalignement

Il s'agit de créer un lien entre les bases de données *texte* et *lexique*. L'interalignement est relatif au *Type de base de données* **du texte** à interaligner. Ainsi tout texte importé suivant ce type sera prêt pour l'interalignement.

- Activez la base de données Texte en cliquant dans sa fenêtre
- Cliquez dans la fenêtre de la base *FriSampl.txt*
- *Projet*, *Type de base de données*
- Sélectionnez *Frisian Text*
- Cliquez sur *Modifier*
- Cliquez sur l'onglet *Interalignement*
- Cliquez sur *Configuration* (rapide)

Définition des marqueurs d'interalignement du texte

Dans le fenêtre *Définition des marqueurs d'interalignement*, il s'agit de définir les marqueurs des lignes que le processus d'interalignement va devoir ajouter dans la base de données texte. La configuration rapide choisie ici définit ces marqueurs de façon standard. Si vous ne voulez rien changer, vous devez néanmoins vous assurer que vous avez bien le bon marqueur pour la ligne de départ. La ligne de départ est celle contenant la phrase à interaligner, c'est donc bien ici la ligne de champ **\t**

Marqueur de texte **t**

```
\t Berne en opgroeid yn Ynje,
```

En dessous de cette ligne existante Shoebox va ajouter une ligne de segmentation dans laquelle sous chaque mot de la ligne t apparaîtra le découpage en morphèmes du mot en dessus. Cette ligne aura comme marqueur de champ **\m**.

Segmentation **m**

```
\t Berne   en opgroeid     yn Ynje,
\m bern -e en op- groei -e yn Ynje
```

En dessous de la ligne de segmentation, Shoebox ajoutera une ligne constituée de la définition de chacun des morphèmes en dessus. Cette ligne aura comme marqueur de champ **\g**

Glose **g**

```
\t Berne      en  opgroeid        yn Ynje,
\m bern -e    en  op- groei -e    yn Ynje
\g bear -PSTP and up- grow  -PSTP in Indonesia
```

En dessous de la ligne de glose Shoebox va rajouter une ligne de catégorie grammaticale pour chaque morphème de la ligne m. Cette information sera dans un champ de marqueur **\p**

Catégorie Gramm. **p**

```
\t Berne      en   opgroeid         yn   Ynje,
\m bern -e    en   op-  groei -e    yn   Ynje
\g bear -PSTP and  up-  grow  -PSTP in   Indonesia
\p V    -Tns  Conj Dir- V     -Tns  Prep N
```

- Cliquez sur OK pour accepter ces paramètres

Configuration rapide des marqueurs lexicaux

La boîte de dialogue *Configuration rapide des marqueurs lexicaux* qui s'ouvre présente dans la fenêtre de gauche les bases de données actuellement ouvertes dans le Projet (texte et lexique). Il s'agit de dire à Shoebox dans quelle(s) base(s) rechercher les informations pour l'annotation du texte. Dans notre exemple, ces informations (morphèmes, gloses, cat. gram...) proviendront de la base de données lexicales, soit *Frirt.dic*.

- Sélectionnez *Frir.dic*
- Cliquez sur *Insérer*

Le segmenteur de Shoebox va rechercher la présence d'affixes dans les mots de la ligne **t** du texte et générer la ligne **m** des découpages en morphèmes. Le processus d'annotation commence à partir de cette ligne **m**.

La définition des morphèmes qui s'affichera sur le ligne \g du texte proviendra du champ *glose* du **lexique** qui a pour marqueur **\g**

Marqueur de glose : **g**

Lexique	Texte
\fri **bern** \ps V \g **bear**	\t Berne en opgroeid yn Ynje, \m **bern** -e en op- groei -e yn Ynje \g **bear** -PSTP and up- grow -PSTP in Indonesia \p V -Tns Conj Dir- V -Tns Prep N

La catégorie grammaticale qui s'affichera sur la ligne \p du texte proviendra du champ *partie du discours* du **lexique**, soit le champ **\ps**

Part. du discours : **ps**

Lexique	Texte
\fri **bern** \ps **V** \g bear	\t Berne en opgroeid yn Ynje, \m **bern** -e en op- groei -e yn Ynje \g bear -PSTP and up- grow -PSTP in Indonesia \p **v** -Tns Conj Dir- V -Tns Prep N

Chaque mot de la ligne de texte est d'abord recherché dans le lexique au niveau du marqueur d'enregistrement . S'il ne s'y trouve pas, il peut être recherché dans un autre champ du lexique, le champ de *forme alternative*. Ce champ (**\a** habituellement), qui peut être répété à l'intérieur d'une même fiche, permet de prendre en compte des formes variantes de l'entrée. Il servira par exemple pour mettre des formes de surface d'un lexème telles qu'elles figurent dans le texte, et dont la structure profonde sera analysée dans le champ suivant (**\u** habituellement) appelé *forme de base*.

Forme alternative : **a**
Forme de base **: u**

- Cliquez sur OK pour accepter les paramètres de l'interalignement

Boîte de dialogue Interalignement

La fenêtre *Interalignement* de la boîte de dialogue *Propriétés du type de Base de Données* récapitule le paramétrage de l'interalignement de la base de données texte. Après une configuration *rapide* de l'interalignement, elle présente 3 lignes :

- Un processus de segmentation qui produit une ligne des découpages \m à partir de chaque mot de la ligne \t du texte
- Un processus d'annotation qui crée une ligne \g des définitions de chaque morphème de la ligne \m
- Un processus d'annotation qui crée une ligne des catégories grammaticales \p de chaque morphème de la ligne \m

Ce paramétrage *rapide* créé à partir du seul renseignement des noms de marqueurs à utiliser dans la base *texte* et dans la base *lexique* peut maintenant être modifié pour répondre plus précisément à un comportement voulu de l'interalignement.

- Sélectionnez le processus à modifier
- Cliquer sur *Modifier*

Segmentation

Le processus de segmentation commence par rechercher chaque mot du texte dans le lexique désigné, généralement au niveau du *champ d'enregistrement* (dans notre exemple **\fri**) et du champ de *forme variante* (**\a**). Si le mot existe dans le lexique, Shoebox recupère en retour le contenu d'un champ particulier **\u** tel que défini dans la configuration rapide, ou à défaut le mot lui même. Pour modifier ce paramétrage standard,

- Cliquez sur le bouton *Lexiques*

Lexique

- Cliquez sur le bouton *Lexiques*
- Sélectionnez, dans la fenêtre des *bases disponibles*, la base de données **lexicales**
- Cliquez sur *Insérer* pour la verser dans la fenêtre des *bases à explorer*
    - Répétez l'opération sur touts les bases à explorer
- Faites de même avec les *champs à explorer* (*entrée lex., forme variante*)
- Sélectionner le *champ à sortir* (*forme sous-jacente*)
- Cliquez sur *OK* pour prendre en compte les modifications

Affixes

Le processus de segmentation ou de découpage morphématique des mots repose ensuite sur le détachement progressif des affixes pour arriver à la racine. Les affixes sont des entrées de lexique comme les mots pleins. Les préfixes et les suffixes sont reconnus comme tels par le segmenteur par le simple fait de comporter un tiret (-) à l'initiale ou en finale. (Les *infixes* comporte un tiret devant et derrière). Par exemple l'entrée **\fri -s** sera considérée comme un suffixe et donc "s" pourra être isolé comme suffixe s'il est trouvé en fin de mot. Par contre **\fri op-** sera un préfixe et donc "op" pourra être isolé à l'initiale d'un mot.

Le principe de la segmentation est de toujours commencer par les segments les plus longs. Ainsi un mot plein trouvé dans le dictionnaire ne sera pas découpé même s'il comporte des segments qui pourrait être interprétés comme des affixes. De même les affixes sont recherchés des plus longs au plus courts. Lorsqu'un segment a été trouvé dans le dictionnaire, il est isolé et ne sera pas découpé plus avant (sauf si une deuxième ligne de segmentation est demandée).

Séquences de recherche

La recherche d'affixes peut être faite selon une préférence particulière suivant que la langue a plutôt tendance à privilégier la préfixation ou la suffixation.

- Préfixes, de préférence
- Suffixes, de préférence
- Pondérer préfixes et suffixes

Infixe devant la racine

Un infixe est un segment qui s'insère à l'intérieur de la racine du mot

- Cochez cette case pour que l'infixe trouvé s'affiche de préférence à gauche de la racine restante. Par défaut l'infixe trouvé sera mis à droite de la racine.

Si la segmentation échoue

Lorsqu'un mot n'a pas été trouvé dans le lexique, Shoebox peut se comporter de différentes façons:

- **Insérer dans le lexique** : cette option ouvre automatiquement une nouvelle fiche dans le lexique pour le mot non trouvé dans le lexique
- **Affiche la marque d'échec** : des astérisques s'afficheront pour un mot ou une racine restante non trouvé(e) dans le lexique
- **Afficher le mot original** : Si le mot ou la racine restante n'est pas trouvée dans le lexique, le mot entier d'origine est recopié simplement sur la ligne de segmentation.
- **Afficher la racine présumée** : Le mot ou la racine restante non trouvé(e) dans le lexique, s'affichera sur la ligne de segmentation.

Délimiteurs de morphèmes

Cette fenêtre permet de ddéfinir la liste des délimiteurs de morphèmes. Ce sont des caractères spéciaux qui sont utilisés par le segmenteur pour reconnaître les affixes dans le lexique. Le tiret (-) est généralement utilisé, mais on peut distinguer des affixes de dérivation (-) et des affixes de flexion (=) par des caractères différents pour rendre l'annotation plus parlante.

Délimiteurs de gloses forcées

Ce sont des caractères appariés, par défaut { }, servant à encadrer une *glose forcée*. Une glose forcée est une information supplementaire ajoutée à la suite du contenu d'un champ dans le lexique, entre deux caractères spéciaux, pour faciliter le processus d'interalignement en lui imposant le contenu de la ligne suivante. Ceci permet de réduire le nombre d'ambiguïtés à résoudre lors de l'interalignement en présence d'homonymie lourde.

Soit par exemple un extrait d'un lexique anglais où il y a deux suffixes *-s* (l'un relatif au nom, l'autre au verbe), et une glose forcée dans l'entrée *man* :

\lx -s \a -es \ps ninfl \ge PL	\lx -s \a -es \ps vinfl \ge 3S	\lx man \a men \u man -s**{PL}** \ps n \ge male_person	<---- glose forcée

Dans le texte, le mot **men** sera découpé automatiquement comme suit

\t	men	
\m	man	-s
\g	male_person	-PL
\p	n	-ninfl

Sans la glose forcée donnée dans le lexique à la ligne \u man -s{PL}, Shoebox aurait été obligé de demander de lever l'ambiguïté entre les deux possibilités PL ou 3S pour le suffixe -s.

Symbole de morphophonologie

Le segmenteur de Shoebox est capable de traiter des cas de transformations morphophonologies d'une racine en présence d'un affixe. Il peut ainsi traiter des cas d'élision ou d'épenthèse à la frontière d'une racine et d'un affixe. Pour cela il faut donner la règle de transformation dans le champ de structure profonde \u **de l'affixe**.

Cela fonctionne ainsi : le champ de la *forme variante* **\a** contient la forme de surface comprenant l'intégralité du suffixe. La champ de *structure profonde* **\u** contient la forme sous-jacente de la partie racine (ou du suffixe précédent) qui est modifiée, suivie du caractère +, puis de la forme sous-jacente du suffixe.

Voici comment les formes des verbes "**try"** et "**tie"** sont découpées, suivant les variantes du suffixe *-ed* définies dans le lexique :

Lexique		Texte					
\lx	-ed	\t	tr**ied**		\t	tie**d**	
\a	-d	\m	try	-ed	\m	tie	-ed
\a	-ied	\g	attempt	-PAST	\g	bind	-PAST
\u	y+ed	\p	v	-vinfl	\p	v	-vinfl
\ps	vinfl						
\ge	PAST						

Notez comment l'interalignement renvoie dans la ligne de segmentation **m** le contenu du marqueur **\u** dans le premier cas et le contenu du marqueur **\lx** dans le second, suivant qu'une forme variante comporte une ligne d'analyse profonde ou non.

Conserver les majuscules et la ponctuation

Ces cases seront cochées dans le cas où on utilise Shoebox pour faire de l'*Adaptation* (traduction entre langues apparentées), c'est-à-dire si l'on cherche à reconstruire des phrases d'une langue 1 vers une langue 2 par des processus de *Segmentation*, *Réarrangement* et *Reconstruction*. Dans ce cas la ligne d'arrivée reflétera la ponctuation de la ligne de départ.

Style segmentation Sh2

Dans la version actuelle de Shoebox, les mots pleins, les racines et les affixes peuvent tous loger dans la même base de données, de même que les règles de transformations morphophonologiques seront définies dans le corps de la fiche lexicale. Dans la version 2 de Shoebox sous DOS, le découpage des mots en morphèmes se faisait en transitant dans un lexique spécialisé de *segmentation* et dans un autre lexique dit d'*affixes conjoints*. En choisissant cette option, de nouveaux boutons apparaissent pour désigner les lexiques à utiliser à cet effet ainsi que les champs à explorer et à renvoyer. A utiliser par les nostalgiques de Shoebox 2.

Respecter les formules de mot

Du fait de la simplicité de base de son segmenteur morphologique, Shoebox peut analyser un même mot de différentes manières et est obligé alors de demander l'arbitrage de l'utilisateur, à travers une fenêtre dite d'*ambiguïtés*. Souvent parmi ces découpages certains ne sont pas valides du fait d'une séquence impossible de morphèmes dans la langue étudiée. Afin de réduire les ambiguïtés qui apparaissent dans la segmentation d'un mot, une option nouvelle permet de définir des séquences valides de morphèmes et par là de supprimer les découpages invalides.

- Cochez la case *Respecter les formules de mot* si vous voulez que Shoebox tienne compte des formules de validité.

Le bouton *Formules* devient alors accessible pour aller définir les formules de mots valides.

Formules

Une formule de mot est composé d'un *symbole* et d'un *modèle*. Pour exprimer par exemple qu'un *Mot* peut consister en une racine nominale *n* (éventuellement suivie par un suffixe nominal *nsuf*) ou en une racine verbale *v* (éventuellement suivie par un suffixe verbal *vsuf*), on écrira :

Symbole:	Mot
Modèles:	n (nsuf)
	v (vsuf)

Soit l'extrait de lexique anglais :

\lx	tiger	\lx	bear	\lx	bear	\lx	-s	\lx	-s
\hm		\hm	1	\hm	2	\hm	1	\hm	2
\ps	*n*	\ps	n	\ps	v	\ps	nsuf	\ps	vsuf
\ge	Felis_Tigris	\e	Ursidae	\e	hold_up	\e	pl	\e	3s

Le mot *tigers* dans le texte s'interalignera automatiquement sans ambiguïté car une fois isolé *-s*, la racine *tiger* est trouvée dans le lexique avec la catégorie grammaticale *n* (nom), la formule précédente impose alors que le suffixe *-s* soit *nsuf* (suffixe nominal) et donc :

\t	tigers	
\m	tiger	-s
\g	Felis_Tigris	pl
\p	n	-nsuf

Sans formule, le mot *bears* produirait à la segmentation deux fenêtres d'ambiguïtés, la première avec *bear* (*n* ou *v*) et la seconde avec le suffixe-*s* (*pl* ou *3s*). Sur les quatre choix possibles, 2 ne sont pourtant pas valides. Avec l'utilisation des formules, il n'y a plus qu'une ambiguïté : soit *bear* est un nom suivi d'un suffixe nominal, soit c'est un verbe suivi d'une suffixe verbal.

\t	bears		\t	bears	
\m	bear	-s	\m	bear	-s
\g	Ursidae	-pl	\g	hold_up	-3s
\p	n	-nsuf	\p	v	-vsuf

Remarque : on peut être plus précis dans la formule en se référant à la *définition* plutôt qu'à la *catégorie grammaticale*

Symbole :	Mot
Modèles	n (pl)
	v (3s)

Voici un exemple de formulation avec des symboles intermédiaires pour rendre les formules plus lisibles

		Commentaire
Symbole	Mot	un "mot" est
Modèles	Nominal	soit "nominal"
	Verbal	soit "verbal"
Symbole	Nominal	un "nominal" est
Modèles	RacNom (pl)	une "racine nominale" éventuellement suivi d'un morphème "pl" (pluriel)
Symbole	RacNom	une "racine nominale" est
Modèles	n	soit un morphème "n" (cat. gram. nom)
	v vnsr	soit un morphème "v" suivi d'un morphème "vnsr" (nominaliseur)
Symbole	Verbal	un "verbal" est
Modèles	v (3s)	soit un morphème "v" éventuellement suivi d'un morphème glosé "3s"
	n vnsr (3s)	soit un morph. "n" suivi d'un morph. "vnsr", évent. suivi d'un morph. "3s".

Par convention

· Les symboles des formules de mot commencent par une majuscule (ex. Mot, Nominal).
· Les données lexicales (gloses) commencent par une minuscule (ex. n et v; pl et 3s).

Voici comment fonctionne la validation du découpage du mot "philisophizes"

\t	philosophizes		
\m	philosophy	-ize	-s
\g	learning	-nvzr	-3s
\p	n	-nsuf	-vsuf

1. philosophizes est-il un Mot ?

Symbole	Mot
Modèles	Nominal
	Verbal

2. philosophizes est-il Verbal?

Symbole	Verbal
Modèles	RacVerbal **(3s)**

- *-s* est-il glosé **3s** ? Oui.

3. philosophize est-il un RacVerbal ?

Symbole	RacVerbal
Modèles	v
	n vnsr

- *philosophy* est-il glosé **n** ? Oui, dans la catégorie grammaticale
- *-ize* est-il glosé **nvzr** ? Oui

Le découpage est valide car

- tous les symboles ont pu être "remplacés par" des données lexicales
- pour chaque morphème du mot, soit le champ \g soit le champ \p contient une donnée qui correspond aux formules — sans que rien ne reste.

Gloses

Les processus d'annotation (glose) d'un morphème consiste à aller le rechercher dans le lexique désigné et à recupérer en retour le contenu d'un champ particulier. Ainsi dans notre exemple *frison* précédent, la ligne de *catégorie grammaticale* **\p** du texte est construite à partir d'une recherche de chaque morphème de la ligne **\m**, dans le lexique désigné *Frirt.dic*, au niveau du *champ à explorer* **\fri**, qui renvoie le contenu du *champ à sortir* **\ps**. Pour contrôler ou modifier ces paramètres :

- Cliquez sur le bouton *Lexiques*

Lexique

- Cliquez sur le bouton *Lexiques*
- Sélectionnez, l'une après l'autre, dans la fenêtre de gauche, les *bases de données* **lexicales** *disponibles*
- Cliquez chaque fois sur *Insérer* pour les verser dans la fenêtre des *bases à explorer*
- Faites de même avec les *champs à explorer* (*entrée lex.*, *forme variante...*)
- Sélectionner le *champ à sortir* (*cat. gram., définition...*)
- Cliquez sur *OK* pour prendre en compte les modifications

Séparateur (de gloses)

L'annotation d'un morphème se fait en recherchant celui-ci dans le lexique, au niveau de l'entrée ou du champ de forme variante, et en renvoyant le contenu d'un champ de glose (*définition, catégorie grammaticale*, *structure sous-jacente*). Un tel champ peut contenir plusieurs gloses séparées par un caractère spécial qui peut être précisé ici - généralement il s'agit du point-virgule. Dans le cas d'un morphème ayant plusieurs gloses possibles, une *fenêtre d'ambiguïtés* présentant les différents choix s'ouvrira lors de l'interalignement pour que l'utilisateur puisse choisir la bonne glose pour le morphème en cours.

Première glose seule

Dans le cas où cette case est cochée, seule la première glose sera prise en compte et il n'y aura donc pas de *fenêtre d'ambiguïtés* lors de l'interalignement.

Appliquer une *table CC* de remplacements (à la sortie)

Des remplacements systématiques peuvent être appliqués, à travers une table de remplacements CC, à la sortie renvoyée par le processus de gloses. En *adaptation* par exemple, on pourra faire passer le mot d'origine à travers une table de correspondances régulières entre certains phonèmes des deux dialectes.

Processus d'Adaptation

L'*Adaptation* dans Shoebox renvoie à la traduction de textes entre des langues apparentées (d'une langue source vers une langue cible). Les fonctionnalités d'*interalignement* de Shoebox peuvent être utilisées à cette fin et ne se limitent pas à une traduction mot à mot, car il peut segmenter les mots des phrases, changer l'ordre des morphèmes et des mots, et reconstruire des mots à partir de ces morphèmes. Quatre processus sont disponibles pour cela :

- le processus *Données* utilisé dans le cas d'un texte déjà interaligné
- le processus *Réarrangement* pour changer l'ordre des morphèmes dans un mot et des mots dans une phrase
- le processus *Glose* pour transposer des morphèmes de la langue source vers la langue cible
- le processus *Reconstruction* qui utilise un jeu de règles phonologiques pour convertir des formes sous-jacentes en formes de surface (dans la langue cible). Il s'agit généralement du dernier processus. Il peut y avoir plusieurs processus successifs de *Reconstruction*