Segmentation sous Toolbox

David Bevan (Trad. Christian Chanard)

1. Introduction

Ce document décrit comment structurer votre base de données pour segmenter sous Shoebox 4. Sous Shoebox 2, il existait deux méthodes pour fournir l'information de segmentation : la base de données des découpages et la segmentation par les affixes conjoints. Avec Shoebox 4, ceci n'est plus nécessaire car le logiciel comporte un segmenteur morphologique qui peut isoler tous les affixes tout en tenant compte des variations morphophonologiques des affixes et des racines. Si vous le voulez, toutes les informations nécessaires au découpage peuvent être maintenues dans votre base de données lexicales.

Ce document utilise des exemples en Anglais avec les marqueurs de champs et le paramétrage de l'interalignement standard Shoebox / MDF

2. Découpage simple: Racines et Affixes

Avec Shoebox, il est indispensable de préciser dans le lexique si un morphème est une racine, un préfixe, un suffixe ou un infixe par l'utilisation correcte de caractères de frontière de morphèmes (normalement le tiret) dans le champ lexème. Un préfixe ***doit*** comporter un tiret final; un suffixe ***doit*** comporter un tiret à l'initiale; une racine ***ne doit pas*** comporter de tiret; un infixe ***doit*** comporter un tiret à l'initiale et en finale. Prenez conscience que ceci vous interdit de mettre dans le champ lexème, un tiret pour une forme lexicale dépendante, (mais vous pouvez utiliser un champ additionnel avec la forme comportant le tiret si nécessaire).

C'est tout ce qu'il faut pour un découpage simple. Par exemple, les entrées lexicales.

\lx	un-	\lx	success	\lx	-ful
\ps	neg	\ps	n	\ps	nadjzr
\ge	OPPOS	\ge	achievement	\ge	ADJZR

produiront l'interalignement correct de unsuccessful:

unsuccessful		
un-	success	-ful
OPPOS-	achievement	-ADJZR
neg-	n	-nadjzr

Des formes contenant des tirets entre morphèmes sont aussi correctement découpées. Par exemple, avec

\lx	non-	\lx	read	\lx	-er
\ps	neg	\ps	v	\ps	vnomzr
\ge	not	\ge	look_at_book	\ge	AGENT

dans le lexique, non-reader sera découpé comme suit:

non-reader		
non-	read	-er
not-	look_at_book	-AGENT
neg-	v	-vnomzr

Notez le progrès par rapport à Shoebox 2 où chaque mot devait être découpé individuellement, et chaque combinaison d'affixes entrée dans le segmenteur d'affixes conjoints.

2.1 Infixes

Un **infixe** est un morphème qui s'insère à l'intérieur d'une racine ou d'un affixe. Certains entendent par infixe un affixe qui n'apparaît qu'entre une racine et un autre affixe; mais dans Shoebox, il s'agirait là d'un simple préfixe ou suffixe, mais pas d'un infixe.

L'Anglais ne connaît pas d'infixes, mais si

\lx	-int-
\ps	intens
\ge	INTENS

était un infixe d'Intensité, alors s**int**uccessful serait interaligné 'correctement' comme

s**int**uccessful		
success	-int-	-ful
achievement	-INTENS-	-ADJZR
n	-intens-	-nadjzr

Dans le découpage, les infixes apparaissent avant ou après la racine ou le radical dans lequel il est trouvé. (Shoebox permet de choisir). Ces exemples affichent l'infixe après. Les infixes peuvent apparaître n'importe où dans le mot, aussi success**int**ul serait interaligné comme:

success**int**ul		
success	-ful	-int-
achievement	-ADJZR	-INTENS-
n	-nadjzr	-intens-

2.2 Racines composées

Les mots constitués de racines composées seront découpés correctement si les deux racines sont dans le lexique. Par exemple, les entrées lexicales suivantes

\lx	black	\lx	bird	\lx	-s
\ps	adj	\ps	n	\ps	ninfl
\ge	dark	\ge	flying_creature	\ge	PL

conduiront à l'interalignement de blackbirds comme suit:

blackbirds			
black	-	bird	-s
dark	-	flying_creature	-PL
adj	-	n	-ninfl

Les composées avec tiret comme sea-green ou mother-in-law seront découpées correctement. Shoebox 4 comporte une option pour autoriser ou non les racines composées.

3. Formes variantes

Si un morphème comporte plus d'une forme de surface, ceci peut être précisé dans un champ de **variante** (alternate form \a):

\lx	a	\lx	telephone
\a	an	\a	phone
\ps	art	\ps	n
\ge	INDEF	\ge	transceiver

La forme reprise par le découpage sera celle du champ \lx:

an	enormous	phone
a	enormous	telephone
INDEF	very_big	transceiver
art	adj	n

Voici des entrées lexicales pour quelques morphèmes comportant des variantes:

\lx	in-	\lx	-s	\lx	not
\a	im-	\a	-es	\a	-n't
\a	il-	\ps	ninfl	\ps	neg
\a	ir-	\ge	PL	\ge	NEG
\ps	neg				
\ge	OPPOS				

et quelques exemples de texte interaligné:

impossible		foxes		faces		haven't	
in-	possible	fox	-s	face	-s	have	-not
OPPOS-	feasible	animal_sp.	-PL	head	-PL	own	NEG
neg-	adj	n	-ninfl	n	-ninfl	v	neg

Remarquez que foxes et faces sont correctement découpés grâce à la présence de fox et face dans le lexique (et pas foxe et fac).

Remarquez aussi que haven't a été analysé comme composé.

Les formes variantes peuvent avoir un type différent du lexème (c.à.d. vous pouvez avoir un suffixe présentant une variante de type racine ou suffixe).

4. Formes sous-jacentes

Si vous souhaitez qu'une forme variante ait sa propre entrée lexicale, vous pouvez utiliser le champ **forme sous-jacente** (underlying form \u). Voici une autre façon de traiter phone. La sortie découpée est la même qu'au-dessus.

\lx	phone	\lx	telephone
\u	telephone	\ps	n
		\ge	transceiver

Vous pouvez entrer le découpage morphématique d'une entrée lexicale dans le champ forme sous-jacente (comme dans l'utilisation de la base de données de segmentation sous Shoebox 2.) Par exemple, voici une façon de traiter la forme supplétive du verbe went.

\lx	went	\lx	go	\lx	-ed
\u	go -ed	\ps	v	\ps	vinfl
		\ge	proceed	\ge	PAST

Vous ***devez*** laisser les espaces entre les morphèmes dans le champ de la forme sous-jacente pour distinguer les racines des prefixes et suffixes. Voici comment cela se découpe :

he	went	
he	go	-ed
3SM	proceed	-PAST
pron	v	-vinfl

Plutôt que d'avoir une entrée lexicale pour chaque forme irrégulière, ces formes peuvent être incluses dans l'entrée principale en utilisant le champ de forme variante ***et*** un champ de forme sous-jacente. Voici une autre façon de traiter went, parmi d'autres formes irrégulières:

\lx	go	\lx	find	\lx	hit
\a	went	\a	found	\u	hit
\u	go -ed	\u	find -ed	\u	hit -ed
\ps	v	\ps	v	\ps	v
\ge	proceed	\ge	locate	\ge	strike

Un champ de forme sous-jacente (\u) est associé au champ \lx ou \a qui le précède. (Si aucun champ \u ne suit un champ \lx ou \a, le contenu du champ \lx est la forme sous-jacente.)

Remarquez que hit est ambigü et par conséquent deux formes sous-jacentes sont données de façon à ce que Shoebox affiche une boîte de dialogue de sélection d'Ambiguïtés.

Voici comment les verbes sont segmentés (avec le choix de la forme Passive pour hit).

went		found		hit	
go	-ed	find	-ed	hit	-ed
proceed	-PAST	locate	-PAST	strike	-PAST
v	-vinfl	v	-vinfl	v	-vinfl

Remarquez que Shoebox est indifférent au fait que l'information de découpage d'une forme soit incluse dans l'entrée principale comme forme variante ou qu'elle constitue une entrée à part enrtière (même dans une autre base de données). Ce choix est le vôtre.

Parfois il est nécessaire d'utiliser un champ de forme sous-jacente pour des séquences d'affixes — un peu comme l'utilisation du segmenteur aux affixes conjoints de Shoebox 2. L'affixe -ability en est un exemple en Anglais:

\lx	-able	\lx	-ity	\lx	-ability
\ps	vadjzr	\ps	anomzr	\u	-able -ity
\ge	ABIL	\ge	NOMZR		

Voici comment readability est découpé:

readability		
read	-able	-ity
look_at_book	-ABIL	-NOMZR
v	-vadjzr	-anomzr

Pour finir, voici quelques exemples de formes sous-jacentes de composés:

\lx	have	\lx	brunch	\lx	lunch	\lx	breakfast
\a	I've	\u	breakfast lunch	\ps	n	\ps	n
\u	I have			\ge	noon_meal	\ge	morning_
\ps	v						meal
\ge	own						

qui se découpe comme suit :

I've		brunch	
I	have	breakfast	lunch
1S	own	morning_meal	noon_meal
pron	v	n	n

4.1 Valeurs imposées — Réduire l'ambiguïté

Lorsqu'un morphème a plus d'un sens, il est parfois utile de spécifier dans la forme sous-jacente lequel est requis de façon que Shoebox n'ait pas à afficher de boîte de dialogue Sélection d'ambiguïtés. Par exemple, deux suffixes anglais ont la même forme -s, aussi dans la forme sous-jacente de la forme irrégulière men, cela vaut la peine de spécifier laquelle est requise. Ceci est possible en insérant la glose entre accolades après le morphème:

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

men se découpe comme suit, sans affichage de boîte de dialogue Sélection d'ambiguïtés.

```
men
man          -s
male_person  -PL
n            -ninfl
```

Pour que les valeurs imposées fonctionnent de cette façon, il faut que la ligne de glose apparaisse, dans la sortie interalignée, immédiatement après la ligne de segmentation (découpage morphématique). Si la catégorie grammaticale apparaissait en-dessus de la ligne de glose alors la valeur imposée aurait dû être spécifiée comme -s{ninfl} ou -s{ninfl}{PL}. Remarquez que des valeurs imposées multiples sont possibles, dans l'ordre des lignes d'interalignement.

Voici un autre exemple, avec l'ambiguïté sur la racine :

```
\lx  bear          \lx  bear
\ps  n             \a   bore
\ge  animal_sp.    \u   bear{carry} -ed
                   \ps  v
                   \ge  carry
```

Avec ces entrées lexicales, bore est découpé comme suit, sans affichage de boîte de dialogue de Sélection d'ambiguïtés. (Naturellement, bore est par lui-même ambigü, puisqu'il signifie également ‘drill a hole’.)

```
bore
bear   -ed
carry  -PAST
v      -vinfl
```

Les valeurs imposées d'un affixe doivent être placées après le tiret, pas avant.

4.2 Découpage direct — Eviter des découpages incorrects

Avec les entrées lexicales suivantes,

```
\lx  hop         \lx  hope        \lx  -s
\ps  v ; n       \ps  v           \a   -es
\ge  jump        \ge  expect      \ps  vinfl
                                  \ge  3S
```

le mot hopes serait mal découpé:

```
 hopes
*hop   -s
*jump  -3S
 v     -vinfl
```

Dans la segmentation, Shoebox résout l'essentiel des ambiguïtés en privilégiant toujours les découpes isolant les affixes les plus longs. Ceci évite à l'utilisateur d'avoir à choisir entre différents découpages — souvent incorrects — à travers de nombreuses boîtes de dialogue de Sélection d'ambiguïtés.

Dans le cas de hopes, cela signifie que c'est -es qui est isolé.

Dans des situations comme celle-ci, pour obtenir le bon découpage, il faut ajouter l'information sur le découpage pour contrecarrer l'analyse incorrecte (une approche différente est présentée plus loin):

```
\lx  hope
\a   hopes
\u   hope -s{3S}
\ps  v
\ge  expect
```

Ceci forcera le découpage suivant de hopes :

```
hopes
hope    -s
expect  -3S
v       -vinfl
```

Voici un autre exemple impliquant les préfixes. Pour éviter le mauvais découpage du mot demisting dans lequel le préfixe demi- est isolé,

demisting		
*demi-	sting	
*half-	hurt	
*num-	n	

il est nécessaire de fournir l'analyse de demist:

\lx	demist	\lx	mist
\u	de- mist{fog}{v}	\ps	n ; v
		\ge	fog

Alors le mot sera découpé correctement :

demisting		
de-	mist	-ing
REVERS-	fog	-PTC
neg-	v	-vinfl

Dans certains cas, il y a plusieurs découpages possibles pour un mot (ou partie de mot) avec des frontières de morphèmes à des endroits différents. Dans ce cas, il est nécessaire de fournir les informations des différents découpages possibles de façon que l'ambiguïté soit reconnue. Voici un exemple :

\lx	do	\lx	doe
\a	does	\a	does
\u	do -s{3S}	\u	doe -s{PL}
\ps	v	\ps	n
\ge	perform	\ge	female_deer

Shoebox présentera une boîte de dialogue de Sélection d'ambiguïtés pour vous permettre de choisir l'analyse appropriée :

does		does	
do	-s	doe	-s
perform	-3S	female_deer	-PL
v	-vinfl	n	-ninfl

5. Morphophonologie

Des variations morphophonologiques peuvent être signalées par les formes variantes et les formes sous-jacentes. Voici comment les règles orthographiques

y → i / __ +ed	*or*	y + ed → ied
y → ie / __ +s	*or*	y + s → ies

s'expriment:

\lx	-ed	\x	-s
\a	-d	\a	-es
\a	-ied	\a	-ies
\u	y+ed	\u	y+s
\ps	vinfl	\ps	vinfl
\ge	PAST	\ge	3S

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

Voici comment les formes des verbes try et tie sont découpées:

tried		tied		tries		ties	
try	-ed	tie	-ed	try	-s	tie	-s
attempt	-PAST	bind	-PAST	attempt	-3S	bind	-3S
v	-vinfl	v	-vinfl	v	-vinfl	v	-vinfl

Voici un autre exemple, pour le redoublement des consonnes devant le suffixef -ed:

\lx	-ed
\a	-d
\a	-pped
\u	p+ed
\ps	vinfl
\ge	PAST

et voici comment hopped est découpé:

hopped	
hop	-ed
jump	-PAST
v	-vinfl

Pour le moment, il n'y a aucun moyen de spécifier des règles générales, et cette information de découpage devrait donc être fournie pour chaque suffixe et chaque consonne susceptible d'être redoublée.

L'approche est équivalente pour les préfixes. Il y a très peu de cas concernant les préfixes anglais, mais si nous supposons que la règle

sp → p/ dis+ __ *ou* dis + sp → disp

telle qu'elle s'illustre par le mot dispirited, soit productive, il faudrait la formuler ainsi :

\lx dis-
\a disp-
\u dis+sp
\ps neg
\ge OPPOS

et dispirited se découperait comme suit:

dispirited		
dis-	spirit	-ed
OPPOS-	vitality	-possessing
neg-	n	-nadjzr

5.1 Sensibilité contextuelle — Eviter les découpages incorrects

Même lorsqu'on n'est pas en présence d'un processus morphophonologique, il peut encore être intéressant, pour réduire le nombre de découpages incorrects, d'utiliser la notation morphophonologique lorsqu'une occurence d'affixe est phonologiquement conditionnée.

Par exemple, avec les allophones du préfixe in- tels que spécifiés ici,

\lx in-
\a im-
\a il-
\a ir-
\ps neg
\ge OPPOS

des mots comme image, imam, iris et iron seraient incorrectement découpés s'ils ne constituaient pas eux-mêmes des entrées lexicales. Comme:

iron	
*in-	on
*OPPOS-	upon
*neg-	prep

Si on utilise la notation morphophonologique pour restreindre la portée des allophones aux occurences suivantes:

\lx in-
\a imb-
\u in+b
\a imm-
\u in+m
\a imp-
\u in+p
\a ill-
\u in+l
\a irr-
\u in+r
\ps neg
\ge OPPOS

alors les découpages intempestifs disparaîtront.

5.2 S'assurer que l'ambiguïté est reconnue

La notation morphophonologique peut être utilisée pour limiter l'effet négatif de l'augmentation du nombre de découpages. Par exemple, ce qui suit spécifie qu'il y a deux découpages possibles de la terminaison des mots en -es. Soit il s'agit d'un simple -s (comme dans pushes) soit (l'information additionnelle) c'est e+s (avec une racine se terminant par -e, comme dans likes):

\lx -s
\a -es
\a -es
\u e+s
\ps vinfl
\ge 3S

Dans les cas (comme hopes) où les racines existent avec et sans -e final (hope et hop), ce que peut voir Shoebox c'est que le découpage est ambigü, mais sans l'information ci-dessus (apparemment redondante) Shoebox choisirait

systématiquement celle qui isole l'affixe le plus long. L'ajoût du découpage supplémentaire oblige Shoebox à afficher la boîte de dialogue de Sélection d'ambiguïtés qui permettra le choix (entre hop -s et hope -s).

Le désavantage de cette approche, c'est que la (fausse) ambiguïté est systématiquement affichée, sauf pour les formes pour lesquelles vous ajoutez une information de découpage spécifique.

D'un autre côté, pour des formes comme does qui sont réellement ambiguës, il n'est pas nécessaire de fournir une information de découpage pour aucune d'elles.

6. Réduplication

Des processus simples de réduplication peuvent être représentés dans Shoebox. L'exemple suivant spécifie une réduplication d'une à trois consonnes suivies d'une voyelle à l'initiale d'une racine, d'un radical ou d'un mot réduplicqué. Pour que Shoebox reconnaisse ceci comme une entrée de réduplication, le champ \lx doit contenir les lettres “dup” quelque part. (Il peut être intéressant de faire apparaître ces lettres en début de champ en sorte que toutes les entrées de réduplication se retrouvent triées ensemble.) Les champ \a sont utilisés pour spécifier le modèle à rechercher par l'utilisation des *variables* définies dans les porpriétés d'encodage de la langue.

```
\lx  dupCV-
\a   [cons][vowel]-
\a   [cons][cons][vowel]-
\a   [cons][cons][cons][vowel]-
\ps  intens
\ge  very
```

L'Anglais n'a pas un usage fréquent de la réduplication, mais si le processus défini ci-dessus était anglais, alors ce qui suit serait découpé correctement comme :

bibig		**stro**strong		**bla**black	
dupCV-	big	dupCV-	strong	dupCV-	black
very-	large	very-	powerful	very-	dark
intens-	adj	intens-	adj	intens-	adj

La Réduplication suffixée peut également être spécifiée, de la même façon que la réduplication de lettres fixes. Par exemple, ce qui suit spécifie la réduplication d'un groupe consonnantique en finale avec un i intermédaire.

```
\lx  -dupiC
\a   -i[cons]
\a   -i[cons][cons]
\a   -i[cons][cons][cons]
\ps  dimin
\ge  a_bit
```

Voici à quoi ressembleraient les découpages :

big**ig**		strong**ing**		black**ick**	
big	-dupiC	strong	-dupiC	black	-dupiC
large	-a_bit	powerful	-a_bit	dark	-a_bit
adj	-dimin	adj	-dimin	adj	-dimin

Il est possible de spécifier le type de réduplication de l'Anglais qui copie une syllabe en remplaçant la voyelle — habituellement par i — comme dans tip-top, tick-tock, criss-cross, flip-flop et wishy-washy. L'entrée serait du genre:

```
\lx  dupCiC
\a   [cons]i[cons]-
\a   [cons]i[cons][cons]-
\a   [cons][cons]i[cons]-
\ps  redup
\ge  intens
```

La Réduplication d'une forme complète serait spécifiée comme suit:

```
\lx  dup
\a   [...]
\ps  redup
\ge  informal
```

En ajoutant des tirets, on peut également spécifier des réduplication totales de préfixes ou suffixes.

Voici un exemple en Anglais de réduplication totale, qui montre également son interaction avec la suffixation et le fait que les formes réduplicquées avec tiret sont correctement découpées.

goody-goody			
dup	-	good	-y
informal	-	nice	-FAMIL
redup	-	adj	-familiar