Choix de tokenisation

Sandrine Ollinger

L'étiquetage en partie du discours de Frantext post 1850 a été réalisé à l'aide de Talismane [4].

Cet instrument offre la possibilité de segmenter les phrases en unités lexicales selon deux méthodes distinctes :

- une segmentation simple : à base de règles uniquement, définies à l'aide d'expressions régulières
- une segmentation à base de patrons : par apprentissage supervisé

Dans tous les cas, la segmentation est réalisée en amont de l'étiquetage.

Au début du projet, nous avons décidé d'opter pour la première de ces méthodes. Nous pensions alors qu'il serait assez simple d'opter pour une segmentation maximale et de n'effectuer qu'un nombre très limité de regroupement. À l'usage, cette position d'une segmentation maximale s'est avérée insatisfaisante et les regroupements, que nous détaillons ci-après, sont en définitive nombreux.

La segmentation est donc réalisée à partir de deux ressources :

- un ensemble prédéfinis de caractères délimiteurs[1] :
  ! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [\ ] ^ _ ` { | } ~ « » – — ― ‘ ” „ ‟ ′ ″ ‹ › ‘ ’ * as well as the various whitespace characters: the space bar, the tab character, the newline, etc.
- une liste ordonnée d'expressions régulières qui définit tout ce qui ne doit pas être segmenté.

[1] Remarque : les caractères délimiteurs sont généralement isolés lors de la segmentation, mais ce n'est pas le cas de l'apostrophe, qui est rattachée au mot qui précède : *Quand ?* est composé de deux segments, *Qu'* est composé d'un seul segment.

Cette méthode ne permet aucune gestion de l'ambiguïté[2]. Une suite de caractères est définie comme étant, ou non, une unité lexicale, indépendamment du contexte.

Il existe donc des regroupements incorrects, qui induisent des erreurs d'étiquetage. Ainsi, dans les exemples suivant, *rendez-vous* est systématiquement regroupé et étiqueté nom commun :

- Vous arrivez en retard à tous vos rendez-vous ? (E115)
- Faites comme lui, rendez-vous sourde à tous les cris, rejoignez la pierre pendant qu'il en est temps. (K244)

1 Définition des unités lexicales composées

À l'issue des travaux réalisés par divers membres de l'équipe, les unités composées ont été définies comme l'ensemble des formes appartenant aux sous-ensembles suivant :

1. Les formes de lemmes composés comportant une espace ou une apostrophe dans Morphalou 3 [1], après double validation manuelle selon 4 critères :
    - Si l'un des mots graphiques du composé n'existe pas seul (*jeun*, *afin*, *fur*) => une seule unité
    - Si le composé est ambigu et que les mots graphiques peuvent se suivre par ailleurs (*bien que*) => plusieurs unités
    - Si le moule syntaxique ne pose pas problème (*à l'aune de*) => plusieurs unités
    - Si le composé est une expression latine (*post scriptum*) => une seule unité.
2. Les formes composées comportant un trait d'union dans Morphalou 3[3].
3. Les formes composées à l'aide de traits d'union commençant par un préfixe répertorié dans GLAWI [2], à l'exception de *très-* et *mon-*, qui ont été identifiés comme étant source de trop d'erreur de regroupement.

[2]La seconde méthode aurait pu nous offrir cette possibilité, mais aurait nécessité une masse de données importante.

[3]Morphalou 3 comporte 10 111 lemmes de ce type. Leur vérification manuelle n'a donc pas été réalisée, à la fois pour des questions de temps et de difficulté à maintenir un jugement uniforme sur un nombre aussi important de cas.

4. Les formes composées à l'aide de traits d'union commençant par un préfixe selon une liste additionnelle obtenue à partir des formes composées de Morphalou 3 et disponible en annexe (page 7).

5. Les nombres composés (*trois cent*, *quatre-vingt-dixième*).

6. Les formes débutant par un trait d'union appartenant à la liste de mots grammaticaux disponible page 8.

7. Les noms propres composés répertoriés dans nos lexiqueslexiques[4].

2 Avertissement

Comme le détaille Véronique Montémont [3], la base de données textuelles Frantext a été initiée dans les années 1960, dans le cadre d'un projet lexicographique. Diverses politiques de mécanographie puis de numérisation se sont succédées au cours de sa constitution. D'une façon générale, nous pouvons distinguer les côtes récentes - dont la numérisation a été réalisée depuis 2000 - des côtes plus anciennes.

récentes : dont le nom commence par la lettre B, E ou R ou se situe entre S300 et S999

anciennes : dont le nom commence par la lettre K, L, M, P, Q ou Z ou se situe entre S001 et S299

Dans les côtes les plus anciennes, une particularité majeur peut être relevée comme étant source d'erreurs : le mauvais encodage de tirets d'incise et de dialogue, couplé à l'absence malheureuse d'espace, provoquant une assimilation à des traits d'union.

- la maman interdira-à juste titre-toute autre tentative. (P455)

Certaines heuristiques ont été mises en œuvre pour limiter ces erreurs, mais elles n'ont pas pu être entièrement éliminées.

[4]Remarque 08/02/2018 : Le regroupement systématique des noms propres composés a été revu, 239 regroupements issus de noms de communes homographes de séquences DET + NC, tels que *Le Passage*, ont été abandonnés.

Références

[1] ATILF. Morphalou, 2015. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[2] Nabil Hathout and Franck Sajous. Wiktionnaire's Wikicode GLAWIfied : a Workable French Machine-Readable Dictionary. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA).

[3] Véronique Montémont. Bâtir une ressource lexicale : l'aventure Frantext. In Jean-Paul Bravard, Natalia Del Fatti, Ioan Negrutiu, and Cristina Vieira, editors, *Les ressources*, Les colloques de l'Institut universitaire de France [2], Saint-Étienne, 2011. Publications de l'Université de Saint-Étienne. Textes issus du colloque du 20e anniversaire de L'Institut universitaire de France, tenu à École nationale supérieure de Lyon du 30 au 31 mai 2011.

[4] Assaf Urieli. *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit.* PhD thesis, Université de Toulouse II le Mirail, 2013.

Annexes

A Lemmes de Morphalou 3 comportant une espace conservés

a b c	à partir de	à vau-l'eau
à brûle-pourpoint	a posteriori	à vrai dire
à califourchon	a priori	ab absurdo
a contrario	à priori	ab origine
à donf	à qui mieux mieux	ad hoc
a fortiori	à quia	ad hominem
à jeun	à tâtons	ad libitum
a latere	a tempera	ad limina
a minima	à tue-tête	ad litem

ad litteram
ad nauseam
ad nutum
ad patres
ad personam
ad rem
ad valorem
ad vitam aeternam
afin de
afin que
alter ego
au demeurant
au fur et à mesure
au visé
auto- destructif
b.a. ba
bel canto
bernard l'ermite
bernard l'hermite
bernard lermite
bernard lhermite
bric-à-brac
buen retiro
c'est à dire
cardio-péricardo-myopexie de gor
carpe diem
casus belli
compact disk
compte rendu
cosa nostra
d'ores et déjà
dare dare
de auditu
de bric et de broc
de commodo
de cujus
de facto
de guingois
de gustibus
de jure
de plano
de profundis
de visu
delirium tremens
deo gratias
deus ex machina
dies irae
don juan
don quichotte
don quichottesque
don quichottisme
don- juanisme
ecce homo
eh ben
eh bien
eh quoi
en catimini
en delà de
en outre
en stand-by
et caetera
ex abrupto
ex aequo
ex cathedra
ex nihilo
ex professo
ex voto
exempli gratia
extra muros
garde champêtre
gratis pro deo
grosso modo
habeas corpus
hasta luego
hic et nunc
honoris causa
id est
in abstracto
in extenso
in extremis
in fine
in pace
in petto
in praesenti
in praesentia
in sillico
in situ
in spiritu
in utero
in vitro
in vivo
intra muros
ipso facto
joint venture
kyrie eleison
lato sensu
lingua franca
loc. cit.
locus citatus
magister dixit
manu militari
mass media
mea culpa
mezza voce
minus habens
missi dominici
modern style
modus operandi
modus vivendi
motu proprio
mutatis mutandis
ne varietur
nec plus ultra
nihil obstat
no man's land
noli me tangere
nota bene
numerus clausus
oh la la
oh là là
one woman show
one-man show

op. cit.
opéra bouffe
operating system
opus citatum
osso buco
over arm stroke
p et t.
p et t
p. et t
parce que
patati et patata
patati patata
pax americana
pax romana
per capita
persona grata
persona non grata
peu ou prou
pique niquer
plan plan
pole- position
post mortem
post scriptum
pourvu que
pretium doloris
prima donna
pro domo
pro forma
punching ball
quant à
quat'z arts
quod erat demonstrandum
res militaris
rock and roll
rock and roller
sainte nitouche
salvé regina
se frotti-frotter
sedia gestatoria
semper virens
sine die
sine qua non
souventes fois
stabat mater
statu quo
stricto sensu
sui generis
tandis que
te deum
terminus a quo
terminus ad quem
terra incognita
terra rossa
terza rima
tout de go
ultima ratio
verbi gratia
vice versa
volens nolens
vomito negro
vox populi
vulgum pecus
white spirit
y compris

B Lemmes de Morphalou 3 comportant une apostrophe conservés

aujourd'hui
baha'i
baha'isme
bernard l'hermite
bernard-l'ermite
bernard-l'hermite
bin's
c'est à dire
c'est-à-dire
c't'
ch'timi
chef-d'oeuvre
chef-d'œuvre
chefs-d'œuvre
ct'
d'arrache-pied
d'emblée
d'ores et déjà
dos-d'âne
entr'aide
entr'aimer
entr'apercevoir
entr'axe
entr'égorgement
entr'égorger
entr'ouvert
entr'ouvrir
feda'i
fedda'i
hors-d'oeuvre
hors-d'œuvre
j'm'en-fichiste
j'm'en-foutisme
je-m'en-fichisme
je-m'en-fichiste
je-m'en-foutisme
je-m'en-foutiste
jourd'hui
jourd'huis

jusqu'au-boutisme
jusqu'au-boutiste
l'on
m'amie
m'as-tu-vu
m'as-tu-vue
m'as-tuvuisme
main-d'oeuvre
main-d'œuvre
mains-d'œuvre
mam'selle
mam'zelle
monte-en-l'air
nid-d'abeilles
nids-d'abeilles
no man's land
pac'que
parc'que
pin's
presqu'ile
presqu'île
presqu'îlette
prud'homal
prud'homie
prud'hommal
prud'homme
qu'en-dira-t-on
quat'z arts
quat'zarts
quelqu'un
quô'c-ngũ'
R'n'B
rock'n'roll
rock'n'roller
sot-l'y-laisse
tape-à-l'oeil
tape-à-l'œil
tout-à-l'égout
traveler's
traveleur's
traveller's
trompe-l'oeil
trompe-l'œil
vau-l'eau

C Liste additionnelle de préfixes

abat-
abdomino-
aborto-
accroche-
acido-
adipo-
adiposo-
adréno-
adynamico-
affectivo-
aides-
aigre-
aigres-
albumino-
alcalino-
aluminico-
alvéolo-
ammoniaco-
ammonio-
analytico-
anarcho-
anhydro-
animo-
ano-
antéro-
apico-
argilo-
arrière-
artério-
assyro-
auriculo-
avant-
bacill-
bacillo-
basi-
basio-
bien-
bilio-
bissau-
bronchio-
bronchiolo-
broncho-
broncholo-
cap-
casse-
cellulo-
centre-
centro-
cérébello-
cérébro-
cervici-
cervico-
cholédoco-
coccy-
cortico-
costo-
cubito-
égypto-
fisco-
fluvio-
franc-
fronto-
glucido-

grand-
grande-
grandes-
grands-
grosso-
ilié-
ilio-
maxillo-
nord-
occipito-
ouest-
physico-
sud-
thyréo-
tibio-
touristico-
toxi-
trachélo-
trachéo-
urétéro-
vésico-
vésiculo-

D Liste de mots grammaticaux

-ce
-ci
-elle
-elles
-en
-il
-ils
-je
-la
-là
-le
-les
-leur
-lui
-m’
-m’en
-même
-mêmes
-moi
-nous
-on
-t-elle
-t-elles
-t-en
-t-il
-t-ils
-t-on
-t-y
-t’
-t’en
-toi
-tu
-vous
-y