Corpus Transcription phonétique des grands corpus littéraires. Les règles du jeu

Michel Bernard

CORPUS

Corpus

5 | 2006
Corpus et stylistique

Transcription phonétique des grands corpus littéraires. Les règles du jeu

Michel Bernard

revues.org

Édition électronique
URL : http://corpus.revues.org/474
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 1 décembre 2006
Pagination : 143-158
ISSN : 1638-9808

Référence électronique
Michel Bernard, « Transcription phonétique des grands corpus littéraires. Les règles du jeu », *Corpus* [En ligne], 5 | 2006, mis en ligne le 19 novembre 2007, consulté le 08 décembre 2016. URL : http://corpus.revues.org/474

Transcription phonétique des grands corpus littéraires. Les règles du jeu

Michel Bernard

1 Les travaux de linguistique quantitative appliquée aux corpus littéraires n'ont porté jusqu'à maintenant, dans leur immense majorité, que sur le signifiant graphique. Les « formes » étudiées par la lexicométrie, par la stylométrie, sont des graphies, définies comme des suites de caractères[1]. Or, le matériau littéraire est au moins autant sonore que graphique. Outre le fait que la transcription du patrimoine oral est un phénomène relativement récent dans notre histoire culturelle, des genres majeurs comme la poésie, genre littéraire *par excellence*, ou le théâtre relèvent essentiellement d'un art de la parole, de la profération publique, ressortissent à des poétiques phoniques[2], auxquelles l'ortho graphe compliquée du français fournit un matériau peu commode. Quiconque a tenté de travailler, par exemple, sur des questions de versification à partir d'un corpus numérisé en mode texte comprendra ce que j'évoque ici.

2 Il est certes possible d'étudier de petits corpus en pratiquant une transcription phonétique. C'est ce que font, pour ne prendre que cet exemple archétypal, Roman Jakobson et Claude Lévi-Strauss dans leur célèbre étude des « Chats » de Baudelaire[3]. Mais il s'agit d'un travail long, minutieux et difficile, impossible à envisager pour des corpus volumineux. La possibilité d'effectuer automatiquement cette tâche ouvre aux chercheurs en littérature et en stylistique la perspective d'études extensives sur le tissu phonique des œuvres littéraires[4].

3 Cette opération, même si elle est effectuée pour l'essentiel par un ordinateur, requiert de la part de l'opérateur une méthodologie rigoureuse pour fournir des données dignes de confiance. Je me propose ici de décrire la procédure que j'estime être la plus sûre, en étayant mon propos sur l'exemple d'une étude[5] sur *La Règle du jeu* de Michel Leiris. J'ai volontairement choisi un texte de prose, pour montrer qu'il n'est pas que la poésie qui puisse bénéficier de cette approche, et un texte moderne, pour faire abstraction des difficultés liées à la prononciation des textes anciens[6].

4 J'ai utilisé deux logiciels mis généreusement à la disposition de tous par des équipes de recherche, que je tiens à remercier publiquement pour cette contribution. Il s'agit de LAIPTTS-SpeechMill (Université de Lausanne[7]) et Mbrola (Faculté Polytechnique de Mons[8]). Ils fonctionnent tous les deux ensemble et permettent respectivement de transcrire un texte graphique en notation phonétique puis de lire ce texte de manière assez réaliste. Des systèmes du même genre (JAWS, par exemple) sont utilisés pour permettre à des aveugles de prendre connaissance d'un texte imprimé à l'aide d'un scanneur, d'un logiciel de reconnaissance de caractères, d'un phonétiseur et d'un synthétiseur vocal.

5 L'opération en elle-même est assez simple : le texte (qui ne requiert pas de préparation particulière) est soumis au logiciel SpeechMill ; il est ensuite lu par une voix synthétique, ce qui permet de vérifier la correction de la reconnaissance[9]. On récupère ensuite un fichier texte qui contient une transcription phonétique. Le code employé n'est pas l'API mais le transcodage est aisé[10]. Au final, on obtient un fichier qui ressemble à celui-ci :

```
də lɑ litERɑtyR kɔ̃sidERe kɔ myn tɔRɔmɑʃi

si lɔ̃ sɑ̃ tjœ̃ ɑ lɑ fRɔ̃tjɛR tRɑse dɑ̃ lə tɑ̃ də
ʃɑkœ̃ də se RəsɔRtisɑ̃ pɑR lɑ legɑlite fRɑ̃sɛz
Rɛgl ɑ kwɑ sɑ nɛsɑ̃s ɑ vuly kil fy sumi sɛ tɑ̃
mil nœf sɑ̃ vœ̃t dø kə lotœR də laʒ dɔm ɑ ɑtɛ̃
sə tuRnɑ̃ də lɑ vi ki lɥi ɑ ɛ̃spiRe lə titR də sɔ̃
livR ɑ̃ mil nœf sɑ̃ vœ̃t dø kɑt Rɑ̃ ɑpRɛ lɑ gɛR ki
lɑvɛ tRɑvɛRse kɔm tɑ̃ dotR gɑRsɔ̃ də sɑ
ʒeneRasjɔ̃ ɑ̃ ni vwɑjɑ̃ gɛR kə də lɔ̃g vɑkɑ̃s
sɥivɑ̃ lɛkspResjɔ̃ də lœ̃ dø
```

6 Il s'agit du début du livre :

DE LA LITTERATURE CONSIDEREE COMME UNE TAUROMACHIE

« Si l'on s'en tient à la frontière tracée dans le temps de chacun de ses ressortissants par la légalité française — règle à quoi sa naissance a voulu qu'il fût soumis — c'est en 1922 que l'auteur de *l'Âge d'homme* a atteint ce tournant de la vie qui lui a inspiré le titre de son livre. En 1922 : quatre ans après la guerre, qu'il avait traversée, comme tant d'autres garçons de sa génération, en n'y voyant guère que de longues vacances, suivant l'expression de l'un d'eux.

L'examen de cette transcription suscitera sans doute des interrogations : le découpage en syntagmes de la chaine verbale n'est pas toujours heureux (problème des liaisons, par exemple), les suspensions liées à la ponctuation ne sont pas restituées, l'emploi du /ɑ/ ou du /ɔ/ peut parfois être contesté (voir par exemple la transcription /tɔRɔ̃jɑʃi/).

7 Mais il s'agit là de difficultés communes à toutes les transcriptions phonétiques et celle-ci, quoique automatique, n'y échappe pas. On pourrait même en ajouter d'autres : doit-on restituer la prononciation probable de l'auteur ? de l'époque ? Comment transcrire les mots d'origine étrangère ? les noms propres[11] ? Comment, dans le texte poétique, rendre compte des diérèses[12], des divers traitements du /ə/ ? La transcription automatique bénéficie cependant d'un avantage certain : elle applique à l'ensemble du texte un traitement uniforme, basé sur des règles constantes, ce qui n'est pas toujours le cas d'un transcripteur humain[13].

8 Le fichier obtenu peut maintenant être étudié à l'aide des logiciels courants de lexicométrie. J'ai utilisé, pour ma part, Lexico3, d'André Salem[14], mais il va de soi que toutes les remarques faites ici s'appliqueraient à d'autres logiciels dotés des mêmes fonctionnalités.

9 La première question de méthode porte sur la définition de l'unité statistique. La notion de « forme graphique » n'a plus de pertinence. Si l'on se contente de définir l'espace et quelques balises comme délimiteurs, on obtient des chaînes de type lotœʀ (« l'auteur »), ou kɔ myn (« comme une »), dont l'index ne présente guère d'intérêt. La comparaison des têtes d'index hiérarchiques de la version graphique et de la version phonétique montre à la fois leur grande redondance et les quelques distorsions introduites par les homophonies et les phénomènes de liaison :

Index des formes orthographiques	Index des formes phonétiques
de	də
la	lɑ
à	e
et	ɑ
d	kə
l	ʒə
que	lə
le	ki
je	sə
un	le
une	dɑ̃
en	mə
qui	de
qu	dy
les	o
il	u
dans	puʀ
me	pɑʀ
des	nə
j	mɛ

10 Notons cependant que la possibilité de repérer des homophones peut avoir des applications littéraires (par exemple, le premier paragraphe comporte les mots « guerre » et guère », qui seraient tous deux décomptés sous la forme /gɛʀ/). Un repérage systématique de tels phénomènes passerait par un codage spécifique du fichier, en constituant des « mots phonétiques » permettant d'établir une correspondance directe entre le mot graphique et sa transcription phonétique (les phonèmes épenthétiques, ceux des liaisons notamment, étant décomptés à part)[15].

11 Mais le traitement de la chaîne parlée est plus facile à envisager en adoptant le phonème comme unité statistique. Pour cela, on introduit un espace entre chaque signe et l'on soumet ce nouveau fichier au logiciel. La « forme » est alors un phonème unique, et le texte tout entier ne contient, au plus, que les 37 phonèmes du français. Voici l'ordre décroissant qu'ils présentent dans notre corpus :

R, ɑ, l, s, e, t, ɛ, i, d, ə, k, m, ɑ̃, p, n, y, ɔ, v, u,
ʒ, ɔ̃, j, z, f, o, b, ɛ̃, w, g, a, ʃ, ø, œ̃, ɥ, œ, ŋ, ɲ

12 On pourrait utiliser cet index et le comparer à ceux d'autres textes, afin de déterminer ses caractéristiques sonores, mais aussi et surtout effectuer des calculs de spécificité à l'intérieur de l'œuvre. Voici, à titre d'illustration, les sonorités spécifiques[16] de chaque chapitre de *L'Âge d'homme* :

Chapitre	Sonorités spécifiques
De la littérature considérée comme une tauromachie	i k t
Je viens d'avoir trente quatre ans	
I-Tragiques	ə œ
II-Antiquités	l b
III-Lucrèce	l ɑ k ø
IV-Judith	ø l
V-La tête d'Holopherne	u b R ɑ
VI-Lucrèce et Judith	m g
VII-Amours d'Holopherne	ɛ m ə n ʒ
VIII-Le radeau de la Méduse	

13 Une comparaison avec le tableau présentant les formes graphiques spécifiques de chaque chapitre permet de retrouver l'origine de certaines spécificités sonores (la fréquence de « frère » dans le chapitre V, par exemple, peut expliquer la fréquence de /R/ dans ce même chapitre) mais les spécificités sonores ont une certaine autonomie et confèrent aux chapitres une coloration particulière.

14 Si l'on veut travailler sur des suites de phonèmes, on utilisera les « segments répétés », tels qu'ils sont calculés par le logiciel[17]. Cet outil, qui sert d'habitude à relever des séquences de mots, permet ici de retrouver des suites de phonèmes, parmi lesquels on repère des mots. Voici un extrait de cet index des segments répétés :

Segments répétés	Fréquence
p e l ɑ	3
p e l ə	4
p e l ɔ̃	2
p e l	10
p e m	3
p e n e	6
p e n e t ʀ	5
p e n e t ʀ e	4
p e n i b l	5
p e n	14
p e p ɑ ʀ	5
p e p	9
p e s i	3
p e s	15
p e s y ʀ œ̃	3
p e t ɛ	2
p e t ʀ i f	2
p e t ʀ	5

15 On y reconnaît des mots entiers (« pénible », « pénétrer ») mais aussi des syllabes, des fragments du discours parlé (les 5 occurrences de /p e p ɑ ʀ/, par exemple, proviennent des passages suivants : « frap**pé par** le tonnerre » (28), « frap**pé par** l'extrême minutie » (71), « Cou**pée par** une loterie » (74), « frap**pé par** le contact » (140) et « préoccu**pé par** ce mythe » (190). Il est ainsi possible de discerner des jeux de mots, calembours, contrepèteries, paronomases, à-peu-près, anagrammes, mais aussi des disséminations thématiques. On pourrait par exemple étudier comment le thème de la « mort » se diffuse dans l'œuvre non seulement à travers les occurrences du mot, de ses flexions, de ses dérivés, de ses synonymes, de son champ lexical mais aussi des 146 occurrences de la chaîne /mɔʀ/, que l'on retrouvera, aussi bien, dans « mortification », « morale » ou « morsure ». L'outil informatique permet ainsi d'étendre à des corpus importants des études qui n'étaient jusqu'à maintenant tentées que sur des textes brefs.

16 Il est ainsi possible, à l'aide de ce dispositif, d'approcher de manière plus scientifique la fameuse question de l'allitération et de l'assonance, qui est une difficulté réelle de la stylistique. En effet, la redondance ne peut être mise en évidence que de manière statistique[18]. On ne peut parler d'allitération, par exemple, que si la concentration d'une consonne est assez forte dans une portion de texte pour détonner par rapport à son utilisation habituelle. Voici des exemples de passages signalés par l'outil informatique[19] :

> Allitérations en /R/ : « la Sainte Vierge, Jeanne d'Arc et Vercingétorix, — espèce de trinité dans laquelle la Vierge et Vercingétorix devaient être plus ou moins en ménage et Jeanne d'Arc représenter, peut-être, leur produit hermaphrodite, vierge guerrière participant des deux et que je serais tenté, pour un peu, de regarder comme préfigurant, grâce à cette double qualité d'être chaste et d'être meurtrière, ces deux images de femmes sanglantes qui sont aujourd'hui dressées dans mon esprit » (58-59), « Très épris de théâtre, et pourvu de remarquables dons comiques, il travailla d'abord pour être acteur, se destinant au Théâtre Français ; » (76), « Oui, je voudrais être enfin compris, je voudrais pleurer dans les bras d'une femme, pleurer sans craindre la raillerie, pleurer sûr d'être consolé ! » (150)

Allitérations en /l/ : « très jeune, la Suisse, la Belgique, la Hollande, l'Angleterre ; plus tard la Rhénanie, l'Égypte, la Grèce, l'Italie et l'Espagne ; » (25), « il y a l'S dont la forme autant que le sifflement me rappelle, non seulement la torsion du corps près de tomber, mais la sinusoïdalité de la lame ; » (29), « Un bond plus violent que les autres et mon frère, perdant l'équilibre, s'effondra en plein sur le vase qui éclata en mille morceaux sur le plancher. Il se releva la tête toute mouillée d'urine. » (117-118)

Allitérations en /s/ : « La seule chose claire que je percevais, c'est le mot « suicide » lui-même, dont j'associais la sonorité avec l'idée d'incendie et la forme serpentine du kriss, et cette association s'est tellement ancrée dans mon esprit qu'aujourd'hui encore je ne puis écrire le mot SUICIDE sans revoir le radjah dans son décor de flammes : » (29), « En voici la substance. » (85), « Dès le moment où le personnage apparaissait sur l'estrade, je prévoyais sa chute (sachant, par l'expérience des précédentes séances, comment le cérémonial se déroulait) et c'était l'attente de cette chute qui le plus m'angoissait. » (110)

Assonances en /i/ : « Je demeurais saisi d'une espèce de vertige en imaginant cette infinie série d'une identique image reproduisant un nombre illimité de fois la même jeune Hollandaise [...] » (34), « C'est lui aussi qui m'a appris qu'il peut y avoir, “plus de poésie dans une chanson à deux sous que dans une tragédie classique”. » (78), « J'obéis à ma vocation et — renonçant aux vagues études scientifiques que j'avais poursuivies jusqu'alors — je quittai le laboratoire de chimie où j'avais fini mon service sans même dire adieu au professeur qui m'y avait accueilli, décidé à consacrer toute mon activité à la littérature. » (181)

Allitérations de dentales /d/, /t/ : « Sur un antique piano droit, dans une salle attenante à celle du comptoir, un homme à mine de vieux grand-père très digne mais un tantinet alcoolique, au poil tout blanc et pauvrement vêtu, tapotait, avec des délicatesses de petite fille, un vaste répertoire de vieilles valses, de mazurkas et de polkas, coupées d'airs d'opérettes, d'hymnes nationaux et de refrains patriotiques » (123-124)

Allitération de fricatives /f/, /v/, /s/, /z/, /ʃ/, /ʒ/ : « J'avais de nouveaux compagnons, avec qui je buvais et philosophais, breuvages, fumée, musique et foule constituant l'excitant mental que nous jugions le plus apte à favoriser l'inspiration. » (188)

Accumulations de nasales : « Je ne m'en plains pas, non plus que je ne m'en vante, ayant une même horreur du genre écrivain à succès que du genre poète méconnu. » (25), « Ces deux dernières courses étaient en tout point admirables ; elles m'ont enthousiasmé, mais ne m'ont rien révélé qui modifiât notablement mon point de vue. » (74)

17 Il faudra évidemment se garder de l'erreur cratyliste qui consiste à donner une signification intrinsèque à des signifiants mais on pourra utiliser ces récurrences pour mettre en parallèle les passages ainsi reliés par une logique phonique. Leiris nous y invite à travers une rêverie proustienne sur le mot « suicide » (29), elle-même marquée par des allitérations curieusement croisées : l'évocation du S est caractérisée par une allitération en /l/, alors que l'allitération en /s/ intervient dans la phrase précédente.

18 Pour résumer le propos qui m'a amené à développer ces quelques exemples, je dirai que la possibilité technique de transcrire phonétiquement de grands corpus nous ouvre un champ d'étude qui est à la fois traditionnel, en ce sens qu'il réinvestit des approches, des notions et des techniques déjà largement employées pour étudier des textes brefs, mais dans le même temps profondément innovant puisqu'il nous est permis de déceler des phénomènes mal connus, à l'échelle d'œuvres ou de corpus d'œuvres. Je ne prendrai pour exemple de ce dernier cas que l'étude de l'univers sonore d'un auteur, d'une œuvre. La statistique peut nous aider à comprendre ce qui est en jeu dans la lecture littéraire et

dans l'impression que nous en retirons, ce qui donne à la poésie de Baudelaire ou à celle d'Aragon leur *timbre* particulier.

19 Il y a là, en effet, un paradoxe ou une limite des études formalistes, auxquelles Jakobson assignait comme première tâche, dès 1935, l'« analyse des aspects phoniques d'une œuvre littéraire »[20]. Cependant, les seuls exemples de ces analyses qu'aient donnés Jakobson ou les critiques de son école ne portent jamais que sur un sonnet, une page de prose, une comptine, des textes brefs où l'on peut, à la main, repérer tous les jeux phoniques. Les effets de « l'œuvre littéraire » sont pourtant davantage procurés par l'ensemble du texte que par le détail d'une page. Quand Gracq écrit « Baudelaire : sensualité liturgique, âme aux échos caverneux, à la sonorité native d'église. La singularité de son timbre poétique tient pour une part non négligeable à ce que les râles réprimés du plaisir, les soupirs amoureux, les privautés d'alcôve, viennent éveiller comme naturellement dans ses vers des résonances de cathédrale. »[21], il ne se réfère pas à tel ou tel poème mais à l'impression auditive qu'a laissée en lui l'œuvre baudelairien[22] dans son entier. Comment pouvoir accéder à de tels phénomènes sans disposer de l'ensemble du matériau sonore d'une œuvre ?

20 La critique littéraire dispose désormais d'outils puissants et accessibles qui lui ouvrent ce champ d'étude. La stylistique et la stylométrie devraient, en particulier, tirer parti de documents qui sont immédiatement accessibles à un spécialiste habitué à travailler sur des transcriptions phonétiques, même si celles-ci n'ont pas habituellement cette ampleur.

21 Il faudra néanmoins, pour populariser ces méthodes, que les outils informatiques soient rendus plus commodes. Les travaux que j'ai pu effectuer ne représentent en effet, pour le moment, qu'un détournement de logiciels qui ne sont pas conçus initialement dans ce but. Voici quelques fonctionnalités qui simplifieraient considérablement la tâche du chercheur littéraire :

- Prise en compte de niveaux de langue anciens, des spécificités du texte poétique, de particularités sociolinguistiques.
- Utilisation standard de l'API par les phonétiseurs et les logiciels de lexicométrie.
- Alignement automatique du texte orthographique et de sa transcription phonétique.

22 En ce qui concerne les outils méthodologiques, il est probable que l'accès à ce nouveau domaine d'étude engendrera un mouvement comparable à celui qui a vu la grammaire de phrase se muer en grammaire de texte. Il faudra en effet forger des concepts et des méthodes capables de rendre compte de phénomènes de récurrence, de symétrie, d'inversion, de contamination à très grande échelle. Dans ce domaine comme dans d'autres secteurs des Etudes Littéraires Assistées par Ordinateur, la machine pose plus de questions qu'elle n'en résout, en donnant accès à des faits qui n'avaient pas été construits comme tels jusqu'alors. Mais elle offre en même temps la possibilité – stimulante – d'y répondre.

BIBLIOGRAPHIE

Beaudouin V. (2000) : *Rythme et rime de l'alexandrin classique. Etude empirique de 80 000 vers du théâtre de Corneille et Racine*, thèse de doctorat, EHESS, 2 vol.

Bechet F., Yvon F., Maurel D. (dir.), Guenthner F. (dir.) (2000) : « Les noms propres en traitement automatique de la parole. Traitement automatique des noms propres », *TAL-Traitement automatique des langues*, 41 (3) : 671-707.

Blain D. R. (1987) : « A Mathematical Model for Alliteration », *Style*, 21 (4) : pp. 607-625.

Capelle J. (1970) : *Manuel programmé d'introduction à la phonétique et à la phonologie du français. 1. Transcription phonétique*, Poitiers : Bureau pour l'enseignement de la langue et de la civilisation française à l'étranger.

Cappeau P. (1997) : « Données erronées : Quelles erreurs commettent les transcripteurs ? », *Recherches sur le français parlé*, 14 : 117-126.

Combettes B., Antoine P., Fresson J. (1970) : *Initiation à la transcription phonétique*, Nancy : Faculté des lettres : I.N.F.A. : C.U.C.E.S. : C.R.D.P

Donnan T. M. (1989) : « Proust "reprit à la musique son bien" : a study in Analogies between Wagnerian and Proustian Composition », *Stanford French Review*, 13 (2-3) : 159-174.

Glidden Hope H. (1994), « "Maulx tant extremes": Homophonic Strategies in Scève's *Délie* », in Nash Jerry C. (éd.), *A Scève Celebration: Délie 1544-1994*, Saratoga : Anma Libriin, 101-113.

Greenberg N.-A. (1980) : « Aspects of Alliteration: A Statistical Study », *Latomus*, 39 (3) : 585-611.

Jakobson R. (1977) : *Huit questions de poétique*, Paris : Editions du Seuil, collection « Points ». Recueil d'articles publié sous la direction de T. Todorov.

Lacure J.-W. (1994) : « A computer study of systematic sound symbolism in classical Japanese verse », *Computers and the humanities*, 28 (6) : 369-374.

Phalèse H. de (2004) : *La Règle du je dans* L'Âge d'homme, Paris : Nizet.

Salem A. (1993) : *Méthodes de la statistique textuelle*, thèse pour le doctorat d'Etat ès-lettres, Université de la Sorbonne Nouvelle-Paris 3, sous la direction de Maurice Tournier, 2 vol.

Tubach J.-P., Boe L.-J. (1985) : *Un Corpus de transcriptions phonétiques : constitution et exploitation statistique*, Paris : Ecole Nationale Supérieure des Télécommunications.

Viprey J.-M. (2002) : *Analyses textuelles et hypertextuelles des* Fleurs du mal, Paris : Champion. Chapitre 5 : « Configurations phonétiques », p. 99 et sqq.

Viprey J.-M., Brunet E. (1997) : « Dynamique du vocabulaire des *Fleurs du mal* », *Travaux de linguistique quantitative*.

Wester M., Kessens J.-M., Cucchiarini C., Strik H. (2001) : « Obtaining phonetic transcriptions : A comparison between expert listeners and a continuous speech recognizer », *Language and Speech*, 44 (3) : 377-403

ANNEXES

Tableau de correspondance entre les codes de SpeechMill et l'API

Symbole phonétique (ASCII 7-bits)	Exemples	API
i	inouit, lit, emis	i
y	utile, vu	y
e	été, créer	e
2 (deux)	deux, bleu	ø
E	vert, treize	ɛ
5 (cinq)	intention, cinq, main	ɛ̃
9 (neuf)	neuf, oeuf	œ
1 (un)	un, parfum	œ̃
a	tabac	a
A	il bat	ɑ
@	ange	ɑ̃
o	oser, galop	o
O	eloge	ɔ
^	on, savon	ɔ̃
u	roue	u
*	premier	ə
%	alpes	ə
j	yeux, paille	j
8 (huit)	huit, lui	ɥ
w	oui, nouer	w
p	père, soupe	p
b	bon, robe	b
m	main, femme	m
f	feu, neuf	f
v	vous, rève	v
t	terre, vite	t
d	dans, aide	d
n	nous, ananas, tonne	n
N	gnocci, agneau, vigne	ɲ
k	carré, lacque	k
g	gare, aigu, bague	g

Symbole phonétique (ASCII 7-bits)	Exemples	API
s	sale, dessous, housse	s
z	zéro, maison, fraise	z
S	chat, achat, tâche	ʃ
Z	gilet, mijoter, fige	ʒ
l	lent, élever, sol	l
R	rue, venir	ʁ
r	rue, venir	r
G	camping (emprunt anglais)	ŋ

NOTES

1. . Voir par exemple la définition donnée par André Salem dans sa thèse (SALEM 93, t. 1, p. 29).

2. . J'emploie cette expression à la suite d'Anne Übersfeld (« Éditorial d'introduction au numéro », *AS/SA* nº3, « Le théâtre », mai 1997, p. 117, < http://www.chass.utoronto.ca/french/as-sa/ASSA-No3/AU1.html>).
3. . « "Les chats" de Charles Baudelaire », *L'Homme*, II, 1962, p. 5-21. Reproduit dans JAKOBSON 1977, p. 163 *et sqq.*
4. . La première entreprise de ce genre, qu'il faut saluer comme telle, est la thèse de Valérie Beaudouin sur le vers classique, qui analyse un corpus de 80 000 vers (BEAUDOUIN 2000).
5. . Cette étude a été publiée dans le volume PHALÈSE 2003.
6. . L'intérêt de Michel Leiris pour les jeux phoniques (*cf. Glossaire, j'y serre mes gloses*) n'est, bien entendu, pas étranger non plus à ce choix.
7. . Il semble que le programme ne puisse plus être téléchargé depuis le site de l'université. Le site MBROLA fournit les références d'autres logiciels « Text-to-Speech » équivalents.
8. . On trouvera toutes les indications utiles sur le site du TCTS Lab de la Faculté Polytechnique de Mons : <http://tcts.fpms.ac.be/synthesis/mbrola.html>.
9. . On peut aussi l'exporter directement dans un fichier son. Pour donner une idée des volumes, le fichier son de la lecture intégrale de *L'äge d'homme* représente 455 Mo (format WAV) et dure 4 h 02 mn.
10. . On trouvera en annexe un tableau de correspondance. J'ai choisi, dans ma transcription, de ne pas utiliser certains signes (distinction entre « schwa obligatoire » et « schwa d'expiration », « "h" aspiré » et « arrêt glottique », etc.)
11. . Sur cette question spécifique, voir BECHET et al. 2000.
12. . Sur ces questions de versification, on consultera BEAUDOUIN 2000, en particulier le chapitre VI.
13. . Voir à ce propos WESTER 2001, qui compare les performances de transcripteurs humains avec celles d'un logiciel, et CAPPEAU 1997, qui analyse les erreurs commises par les transcripteurs.
14. . On trouvera toutes les indications utiles à l'adresse < http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/lexico3.htm>.
15. . On trouvera un exemple de telles études sur les homophones dans GLIDDEN 1994.
16. . Spécificités inférieures à 1E-3. Le texte de la transcription phonétique a été balisé de manière classique, comme le requiert Lexico3.
17. . Pour une présentation complète de la notion et du traitement, voir SALEM 1993, p. 154 et sqq.
18. . Approche tentée par GREENBERG 1980 et BLAIN 1987. On trouvera une étude informatisée dans LACURE 1994.
19. . Une des fonctions de Lexico3 permet de découper le texte en portions délimitées par un caractère (j'ai choisi le point et autres ponctuations marquant habituellement la fin d'une phrase) et de calculer la spécificité d'une forme ou d'une liste de formes dans chaque segment. On repère ensuite, grâce à un code couleur, les spécificités les plus fortes. Les paginations renvoient à l'édition Folio Gallimard.
20. . « La dominante », conférence de 1935 dont le texte est reproduit dans JAKOBSON 1977, p. 77.
21. . *Carnets du grand chemin*, Corti, 1992, p. 264.
22. . On lira, sur l'aspect phonique des *Fleurs du mal*, les travaux de Jean-Marie Viprey : VIPREY 1997 et 2002.

RÉSUMÉS

Transcription phonétique des grands corpus littéraires. Les règles du jeu
Le perfectionnement des phonétiseurs permet aujourd'hui d'envisager la transcription phonétique de grands ensembles textuels et, par conséquent, de doter la stylométrie de nouvelles capacités dans le domaine de l'analyse des effets sonores. Cet article montre, sur l'exemple d'une étude de *La Règle du jeu* de Michel Leiris menée à l'aide de LAIPTTS-SpeechMill (Université de Lausanne) et Mbrola (Faculté Polytechnique de Mons), quelles procédures et quelles précautions on se doit d'associer à de telles entreprises. Le repérage et l'étude des allitérations et des assonances, des homophones, la détermination de la tonalité phonique d'une partie du texte montrent tout à la fois les perspectives et les limites de ce type d'approche du texte littéraire.

Phonetic Transcription of Large Literary Corpora. The Rules of the Game
The improvement of the phonetizers makes it possible today to consider the phonetic transcription of large textual sets and, consequently, to give stylometry new capacities in the field of the analysis of the sound effects. This article shows, on the example of a study of *The Rule of the Game* of Michel Leiris carried out using LAIPTTS-SpeechMill (University of Lausanne) and Mbrola (Polytechnic Faculty of Mons), which procedures and which precautions should be observed. The location and the study of alliterations and assonances, the homophones, the determination of the phonic tonality of parts of the text show both the prospects and the limits of this type of approach of the literary text.

AUTEUR

MICHEL BERNARD

« Centre de recherche Hubert de Phalèse », EA 3423, Paris 3