Pour toute information ou commande :
BIL
Université de Lausanne
Faculté des Lettres
I.L.S.L, BFSH2
CH1015 - Lausanne

BULLETIN DE L'INSTITUT DE LINGUISTIQUE

SCIENCES DU LANGAGE
DE LA FACULTE DES LETTRES
DE L'UNIVERSITE DE LAUSANNE

AVERTISSEMENT

La publication du BIL a connu un temps d'arrêt qu'explique principalement une réflexion engagée par le comité de rédaction sur le contenu du bulletin comme sur son mode de diffusion. Cette réflexion est aujourd'hui terminée. Aussi que l'auteur de la contribution présentée dans ce numéro nous pardonne le retard qu'aura pris cette dernière

Le comité de rédaction

ISSN 1023-134X

Du k-gramme au mot

Variation sur un thème distributionnaliste

I n n o v a f e r t a n i m u s m u t a t a s d i c e r e f o r m a s c o r p o r a ; d i , c o e p t i s (n a m v o s m u t a s t i s e t i l l a s) a d s p i r a t e m e i s p r i m a q u e a b o r i g i n e m u n d i a d m e a p e r p e t u u m d e d u c i t e t e m p o r a c arm en! n u l l u s a d h u c m u n do pr ae b eb at l u m i n a T i t a n , n e c n ov a c r e s c e n d o r e p ar a b a t c o r n u a P h oe b e , n e c c i r c u m fu so pe n d e b a t in a e r e t e l l us p o n d e r i b u s lib r a t a s u i s , n ec b r a cch i a l o n g o m a r gin e t e r r a ru m p o rr e x e r a t A m ph i t r i t e ; u t q u e e r a t e t t e l l u s i l l i c et po n t us et a e r , s i c e r a t i n s t a bi l i s t e l l us , i n n a b i l i s u n da , l u c i s e g e n s a e r ; n ul l i s ua fo r m a ma n e b a t , o bst a b a t qu e a l i i s a l i u d , q u i a c o r p o r e i n u n o f r i g id a p ug n ab a nt ca li d i s , u m e n t i a s i c c i s, m o l l i a c u m d u r i s, si n e p o n d e r e , h a b e n t i a p o n d u s . H an c d e u s e t m e l i o r l it em n a t u r a d i r e mi t . n a m c a el o t e r r as e t t e r r i s a bs ci d i t u n d a s e t l i q uid u m spis so s e c r evit a b a ere c a e l u m. q u a e po s t q u a m e v ol vi t c a ec oq ue e xemit a c e r vo , d i s s o c i a t a l oci s c on c o rd i p a c e l i g a v i t: ig n ea c on v e x i v i s et s in e p o n de re cael i e m ic ui t s um maque lo c u m si b i f ec i t i n a rc e; p ro x i m u s e s t ae r i lli l e vi t a t e l o co q u e; d e n s i o r h i s t e llu s el em e nta q ue grandi a t r ax i t et pressa e st g ra vit a te s ua; c i r cumf l u us umor u l t ima po s s ed i t so l i dum qu e c oer c ui t orbem. Si c u bi dis po s i t a m qu i s quis f uit i l l e d e or u m c o n ge rie m s ec ui t s ec t amq ue i n m e m b r a c o e g i t, p r i n c i p io ter r a m , ne no n a e q uali s ab omn i p ar t e fo r e t, m a gni s p e ci em glo m e r avi t i n orb is. tu m f r et a dif fun d i ra p idisq ue t u me s cere vent i s i ussi t et a mb it ae c ircu m d are l i t o r a terrae ; addi di t et fon tes e t st a g na i nm e n sa l acus q u e f lu m in aqu e ob l i qu is c inx i t d ec l iv i a rip is , qua e, d iv er sa loci s , pa rti m s orb e nt ur a b ip s a, in m a re perve ni un t partim ca mpo q ue re cep ta l ib e r i ori s aquae p r o r ip i s l itora pu l sa nt. quarum qu ae m ed ia e s t, no n est habita bili s a est u ; n ix tegit a l ta d uas; tot id e m int e r utra mque l o cav it t emp eri emque de dit mix ta c u m f rigor e f lamma. I nmine t his a er , q ui, q u an to est po n d ere terr ae p on d us a quae levius , tan to e st onerosi or ign i. illic et nebula s , i ll i c c on si ste r e nubes iu s sit et humanas motura ton it r ua mentes et cum f ul minibus fac ien t e s f ul gu ra ventos . His qu oque n o n passim mundi fabricator habend um aera permisit; vi x nunc obs istitur il l is, cum sua quis que rega t diverso f lamina tractu, q uin l anie nt mu ndum ; t ant a est discordia fr a t r um. Eurus a d Auroram Nab atae aque regna recess it Pe rsi daqu e et rad iis iuga subdita ma t uti nis; vesper et occidu o q uae litora so le tep escun t, prox ima sunt Zephyro; S c y th i am septemque trion es horrife r i nvasit Boreas; contraria tellus nubibus adsiduis pluviaque ma de scit ab Austro. haec super inposuit liq ui dum et gr av itate c arente m aethe ra ne c quicquam te rrenae faecis habentem. Sanctius his animal mentisque capacius altae deerat adhuc et quod dominari in cetera posset: natus homo est, sive hunc divino semine fecit ille opifex rerum, mundi melioris origo, sive recens tellus seductaque nuper ab alto aethere cognati retinebat semina caeli. ver erat aeternum, placidique tepentibus auris mulcebant zephyri natos sine semine f1ores; mox etiam fruges tellus inarata ferebat, nec renovatus ager gravidis canebat aristis; f1umina iam lactis, iam f1umina nectaris ibant, f1av

Aris Xanthos
Université de Lausanne

Table des matières

Présentation

Prof. Remi Jolivet
(Université de Lausanne)

Ce numéro publie le mémoire de licence d'Aris Xanthos, assistant à la section de linguistique. Ce travail est consacré à la reprise et au développement d'une question classique, celle de la segmentation de la chaîne en unités significatives sur la base de statistiques des phonèmes.

C'est l'occasion d'un retour aux textes originaux de Harris qui n'est pas sans surprise puisqu'Aris Xanthos montre quels gauchissements a subi la formulation initiale. Ce qui justifie la pertinence et fonde l'originalité des développement qu'il en propose.

Ces développements sont soumis à l'expérimentation sur deux langues morphologiquement fort différentes, l'anglais et le latin.

Ce travail s'inscrit dans un domaine de recherche qui retrouve, depuis quelques années, les faveurs des linguistes. Ils ne peuvent guère y travailler seuls et le concours des mathématiciens y est essentiel. Ce n'est pas diminuer les mérites - qui sont grands - d'Aris Xanthos que de remercier le professeur François Bavaud, de la section d'Informatique et Méthodes Mathématiques, pour son implication dans le suivi de cette recherche.

Il reste à souhaiter à Aris Xanthos que s'ouvre devant lui une riche carrière de chercheur. Il l'a brillamment commencée.

Remi Jolivet

Remerciements

Merci à Remi Jolivet pour m'avoir toujours soutenu dans ce genre d'exotisme intellectuel, sans jamais me laisser partir en dérive. Merci à François Bavaud pour m'avoir accueilli comme un interlocuteur plutôt qu'un élève, et m'avoir pourtant dispensé ses enseignements avec une générosité incomparable. Sans leur bienveillante attention, ce mémoire serait encore plus éloigné de ce que j'aurais voulu qu'il soit.

Merci à tous les membres et amis de la section de linguistique de l'UNIL pour toutes sortes de raisons, qui vont de leur bonjour (plus ou moins) matinal à leur aide très concrète pour la production de données ou la recherche bibliographique. Merci en particulier à Marianne Kilani-Schoch, Pascal Singy, Yves Erard, Laurent Gajo et Christophe Pythoud. Merci également à Martin Forst, notamment pour être mon unique contemporain d'exotisme intellectuel. Merci également à Michael Brent (John Hopkins University) pour avoir aimablement satisfait mes demandes d'informations quant à ses articles.

Merci à tous mes amis et amies pour leur compréhension face à mon isolement - qui fut nécessaire à la rédaction de ce mémoire, et dans l'ensemble pour leurs encouragements à tenir bon et leurs promesses de lendemains qui chantent.

Merci par-dessus tout à mon père Dimitris pour son aide et son soutien quotidien pendant ces mois de rédaction et aussi loin que je me souvienne, et à mon frère Nicolas pour m'avoir appris les règles de ce jeu de langage - pour m'avoir tout bonnement appris l'université. Ce mémoire leur est dédié, ainsi qu'à ma mère infiniment regrettée Monique.

Yverdon, février 2001

Conventions de notation

0.1 Transcriptions

De façon générale, les segments seront transcrits entre barres de fraction en notation phonologique (respectant les conventions de l'API[1] sauf mention contraire) ou orthographique; ce cas sera toujours signalé par l'italique:

/fɔrm/ *forme*

Il sera parfois nécessaire de donner une précision phonétique, ce que l'on indiquera par la mise entre crochets de l'expression concernée:

[fɔnetik]

A l'oral comme à l'écrit, l'espace désignera une frontière de mots; les symboles '=' et '_' marqueront respectivement les frontières *manquées* (non détectées à tort) et les *fausses alarmes* (frontières détectées à tort):

une forme=fau_tive

A quelques occasions, les frontières de morphes à l'intérieur d'un mot seront explicitées par l'usage du tiret:

/diviz-ibl/ *divis-ible.*

1. Le site de la section de linguistique de l'UNIL propose une introduction aux symboles de l'Alphabet Phonétique International (API) et à la phonétique articulatoire: http://www.unil.ch/ling.

0.2 Notations logiques et mathématiques:

$:=$	défini comme
$=$	égal à
$\neq$	différent de
$\approx$	équivalent à (sous une hypothèse donnée)
$<(\leq)$	plus petit que (ou égal à)
$>(\geq)$	plus grand que (ou égal à)
$\cap$	intersection
$\in(\notin)$	(non) appartenance
$(\forall x)P(x)$	«La propriété P est vraie pour tout x»
$\{x : P(x)\}$	ensemble des x vérifiant P
$\|x\|$	valeur absolue de x
$A := \{a_1,...,a_m\}$	alphabet de m symboles
A^k	alphabet d'ordre k ou ensemble de k-grammes
$\|A\|$	cardinal ou taille de A
$\overrightarrow{c}$	coefficient de transition
$\overrightarrow{c}(w), \overleftarrow{c}(w)$	nombre de successeurs / prédécesseurs de $w \in A^k$
$d(\sigma,i)$	variable de décision pour le test de la présence d'une frontière morphologique après le i-ème symbole de σ
d_k	entropie résiduelle d'ordre k
ε	chaîne vide
$E_{a\in A}$	espérance mathématique (ou moyenne théorique) sur les éléments d'un ensemble
$f(x_1,\ldots,x_n)$	fonction générique admettant n arguments
h_∞	taux d'entropie
$\overrightarrow{h}(w), \overleftarrow{h}(w)$	entropie conditionnelle (standard et inverse) étant donné $w \in A^k$
$\overline{h}(w,w')$	moyenne des entropies conditionnelles standard et inverse étant donné le contexte précédent $w \in A^k$ et le contexte suivant $w' \in A^k$

$\overline{h}(w,w',\overrightarrow{c})$	idem, où $\overrightarrow{c}$ pondère l'entropie conditionnelle standard et $1-\overrightarrow{c}$ l'entropie conditionnelle inverse
h_k	entropie conditionnelle d'ordre k
H_k	entropie d'ordre k
H_0,H_1	hypothèses opposées dans le cadre d'un test
$H(p)$	entropie sur la distribution p
$i(a)$,$i(w)$	information associée à $a \in A$ / $w \in A^k$
$i(w \rightarrow a)$,$i(a \leftarrow w)$	information conditionnelle associée à la transition $w \rightarrow a$ / précédence $a \leftarrow w$
$I(w \rightarrow a)$,$I(a \leftarrow w)$	information mutuelle associée à la transition $w \rightarrow a$ / précédence $a \leftarrow w$
$\overrightarrow{I}(w)$, $\overleftarrow{I}(w)$	information mutuelle entre w et ses successeurs/prédecesseurs
$\overrightarrow{I_k}$	information mutuelle d'ordre k
k	ordre (d'un processus, d'une mesure, etc.)
$K(p \parallel p')$	divergence de Kullback-Leibler ou entropie relative entre les distributions p et p'
log	logarithme binaire
$L := \{w_1,\ldots,w_l\}$	lexique de l mots
m	taille de l'alphabet A
$max(x_1,\ldots,x_n)$	le plus élevé parmi $x_1,\ldots,x_n$
n	longueur d'une séquence de symboles
$n(a)$,$n(w)$	nombre d'occurrences de $a \in A$ / de $w \in A^k$ dans un corpus
$p := p_1,\ldots,p_m$	distribution de probabilités sur A
$\widetilde{p}$	distribution initiale (des symboles en début de mot)
p^k	distribution de k-grammes (non conditionnelle)
p^w	distribution conditionnelle (des symboles en début de mot)
$p(a)$,$p(w)$	probabilité de $a \in A$ / de $w \in A^k$
$\widetilde{p}(a)$	probabilité initiale de $a \in A$

$p(x\|y)$	probabilité conditionnelle de x étant donné y
$p(w \rightarrow a)$,$p(a \leftarrow w)$	probabilité de transition de w vers a / de précédence de w par a
$\overrightarrow{P}, \overleftarrow{P}$	matrice de transition / précédence
$\overrightarrow{P}_{wa}, \overleftarrow{P}_{wa}$	composante d'une matrice de transition / précédence
$\sigma := s_1^n := s_1 ... s_n$	séquence de longueur n
Σ	somme
$T(\sigma,i)$	seuil pour le test de la présence d'une frontière après le i-ème symbole de σ
$\overrightarrow{v}(w), \overleftarrow{v}(w)$	variété conditionnelle (standard et inverse) étant donné $w \in A^k$
$w \rightarrow a$,$a \leftarrow w$	transition de w vers a / précédence de w par a
$w \in A^k$	k-gramme (combinaison de k symboles)
$w \in L$	mot

Avant-propos

Dans l'article «From Phoneme to Morpheme» (Harris 1955a[2]), Zellig S. Harris décrit une procédure permettant de segmenter un énoncé (transcrit phonologiquement) en mots et en morphes sur la seule base des propriétés distributionnelles des séquences de phonèmes observées dans la langue considérée. La version *la plus simple* de sa méthode est fondée sur l'observation du *nombre de successeurs* (ou nombre de phonèmes différents susceptibles d'apparaître) en chaque point de l'énoncé; en particulier, Harris indique que la plupart des frontières morphologiques[3] sont situées à des points où la courbe du nombre de successeurs marque un pic et, inversement, que la plupart des pics de cette courbe correspondent à des frontières morphologiques; la segmentation morphologique peut donc être accomplie par simple *seuillage* du nombre de successeurs.

La mise en œuvre de cette approche implique qu'on dispose d'un modèle de la langue si détaillé qu'il est irréaliste de chercher à l'estimer à partir d'un corpus fini d'énoncés (Harris 1955a, p.38). Mais dans un article ultérieur, Harris rapporte les résultats encourageants d'une application informatique de sa méthode à la détection des frontières de morphes *à l'intérieur de mots*[4] (Harris 1967). Dans cette seconde expérience, les paramètres du modèle (c'est-à-dire *les* nombres de successeurs) sont estimés à partir d'un corpus anglais écrit - un dictionnaire pour être précis.

Même si Harris souligne que, dans un cas comme dans l'autre, les résultats de la procédure ne sont pas *exacts*, et qu'elle ne constitue qu'une *étape* de l'analyse morphologique proprement dite, on y a souvent vu la promesse d'une automatisation totale du processus de découverte des signifiants de la langue - enjeu

2. Cet article et celui de 1967 ont été réimprimés dans le recueil «Papers in Structural and Transformational Linguistics» (Harris 1970); dans la suite, nous y référerons par l'année de leur première publication mais citerons en fait la pagination des textes réédités.

3. Nous parlerons fréquemment de frontière (ou de segmentation) *morphologique* lorsque notre propos ne justifie pas la distinction entre morphe et mot.

4. Voir à ce propos la citation faite dans la note 6 (page xx).

considérable pour un certain nombre des *tâches* qui dirigent l'évolution de la linguistique computationnelle. Sous l'impulsion de cette vision exaltée, les chercheurs se sont emparés du nombre de successeurs pour lui faire subir diverses mutations, souvent suggérées par des contraintes liées à la structure des données traitées; de référence en référence, par l'effet d'un genre de *téléphone arabe* bibliographique, les hypothèses initiales de Harris ont été progressivement élidées, alors même qu'on en introduisait, plus ou moins explicitement, de nouvelles. Près d'un demi-siècle plus tard, cette progression aboutit à des formulations telles que:

> If one monitors the instantaneous entropy of a language model as it scans across an English text, one generally finds that regions of high entropy correspond with word boundaries. [...] Segmentation is a matter of *chunking* the data whenever the instantaneous entropy exceeds some threshold value. (Hutchens & Alder 1998, soulignés par les auteurs)

Dans cette version, on s'efforce de détecter les frontières de mots dans un texte littéraire anglais, aux points où l'entropie conditionnelle est élevée; mais la différence fondamentale entre cette approche et celle de Harris et que les auteurs ne tiennent pas compte des frontières d'énoncés, et donc que leur système opère sur un corpus non segmenté. Le fait que ces problématiques essentiellement linguistiques ne fassent pas l'objet d'une réflexion spécifique dans la référence citée n'est pas critiquable, puisque les auteurs travaillent dans le domaine de la compression textuelle et que ces distinctions ne sont pas pertinentes dans le cadre de leur recherche. Mais il est clair que ce genre de référence implicite, indirecte et abrégée contribue à déformer, petit à petit, l'hypothèse initiale de Harris.

Car dans son approche, l'élévation du nombre de successeurs aux frontières morphologiques n'est qu'un cas particulier d'une relation plus générale:

> If we segment the utterance at those points where the sequence looks as it does at utterance beginning or end, we get a segmentation which agrees very well with word and morpheme boundaries for that utterance. (Harris 1955a, p.62)

Dans cette perspective, le critère de base pour l'insertion d'une frontière morphologique après une séquence de symboles n'est pas la *diversité* des successeurs de cette séquence, mais la *similarité* entre la distribution de ses successeurs et celle des phonèmes en début d'énoncé[5], dont on admet qu'elle est une bonne approximation de la distribution des phonèmes en début de mot[6]. Cette similarité peut prendre des aspects différents selon les langues considérées; dans le

5. c'est-à-dire la distribution des successeurs d'une fin d'énoncé

6. Pour apprécier la portée explicative de la méthode de Harris, il peut être éclairant de la concevoir comme une forme d'*apprentissage par généralisation*, en l'occurrence généralisation

cas de l'anglais et de nombreuses autres langues, on peut raisonnablement supposer que le nombre de phonèmes pouvant apparaître immédiatement après une frontière d'énoncé devrait être plus élevé qu'aux autres positions (de l'énoncé), et que la connaissance des symboles spécifiques susceptibles d'apparaître en un point n'est pas indispensable pour prendre la décision d'insérer ou non une frontière. Ce sont ces hypothèses qui justifient l'utilisation du nombre de successeurs - ou d'une autre mesure de diversité - en lieu et place d'un véritable indice de similarité.

Dans une langue où l'une au moins de ces hypothèses n'est pas vérifiée[7], la diversité ne peut pas se substituer à la similarité, et l'application de la procédure implique qu'on dispose non seulement d'un modèle explicite de la distribution des successeurs pour la langue considérée, mais aussi de la distribution des phonèmes en début d'énoncé. Dans de bonnes conditions, les deux modèles peuvent être extraits du corpus même que l'on s'efforce de segmenter; mais que se passe-t-il si le second modèle n'est pas connu, par exemple parce que les frontières d'énoncés ne sont pas transcrites dans le corpus ? A partir de quels exemples positifs peut-on alors effectuer une *généralisation* (voir note 6) ? Et si les deux modèles sont correctement estimés, comment mesurer leur différence d'une façon qui, contrairement au nombre de successeurs, tienne compte des symboles spécifiques considérés[8] ? Ce sont les préoccupations qui sont au cœur de ce mémoire.

Avant de formuler nos propositions à ce sujet, nous devons mentionner un problème annexe liés à l'utilisation de corpus dépourvus de segmentation *a priori*. Dans ce cas, si la séquence est suffisamment longue pour permettre d'estimer correctement la probabilité des symboles et séquences de k symboles pour k petit (de l'ordre de trois ou quatre), elle ne *peut pas* l'être assez pour estimer correctement le nombre de successeurs sur toute sa longueur. C'est là un trait général de la modélisation distributionnelle des séquences de symboles, qui conduit

---

du seul indice univoque de rupture morphologique universellement attesté dans les langues naturelles: la pause de fin d'énoncé. Cette perspective rend possible d'envisager l'application de la méthode à des données autres que linguistiques, pourvu qu'elles incluent une première forme de segmentation dont on a des raisons de penser qu'elle repose sur des critères susceptibles d'être généralisés avec profit au cas d'une segmentation plus fine. Notons que, selon Harris, ce sont bien les frontières de *mots* qui résultent de la généralisation des frontières d'énoncés, les frontières de morphes résultant de celle des frontières de mots: «[...] the points where sequences that characterize utterance boundaries appear within an utterance [...] correlate in general with word boundaries. Other periodicities between word boundaries lead to subsidiary segmentation which correlate with morpheme boundaries» (Harris 1955, p.60).

7. et dans le cas où la transcription utilisée inclut des symboles marquant explicitement les frontières de morphes (voir Harris 1955, p.65, note 5)

8. et non seulement de l'information «abstraite» de l'occurrence de types non identifiés

souvent à l'expression des hypothèses (dites *markoviennes*) de *conditionnement limité* et de *stationnarité*: dans le cas de la méthode de Harris, elles reviennent à considérer le nombre de phonèmes pouvant suivre non plus un début d'énoncé de taille variable mais une séquence de k phonèmes, où la constante k dénote l'*ordre* du modèle. On réduit ainsi considérablement le nombre de paramètres à estimer, mais le coût de cette simplification est une surévaluation de la variété conditionnelle[9]; il est vraisemblable, en effet, que le nombre de symboles différents pouvant suivre une séquence de k symboles dans un corpus fini soit supérieur au nombre de symboles différents pouvant suivre une séquence de $k+l$ symboles ($l > 0$) dans le même corpus.

A ce biais statistique s'en ajoute un second, d'ordre linguistique cette fois-ci: l'articulation en énoncés est une propriété intrinsèque de la communication verbale, et l'absence des frontières d'énoncés dans un corpus entraîne des conséquences sérieuses sur sa fiabilité. Par exemple, si l'on retire ces frontières d'un corpus français dont l'un des énoncés se termine par *nous* et le suivant commence par *vais*, le corpus résultant exhibera la cooccurrence fâcheuse *nous vais*. La conjugaison de ces deux types d'effets aboutit immanquablement à une segmentation *approximative*, que l'on s'efforcera de corriger au paramétrage du modèle. A cet égard, le principal intérêt d'un article comme celui de Hutchens & Alder (1998) est de démontrer[10] que, dans une langue comme l'anglais où la diversité est une estimation valable de la similarité avec la distribution des phonèmes en début d'énoncé, l'utilisation d'un indicateur *tenant compte des fréquences*[11], tel que l'entropie conditionnelle (*instantanée*, dans la terminologie des auteurs cités), suffit à induire la position des frontières de mots - même s'il n'est conditionné que sur un contexte de taille fixe de l'ordre de quelques symboles successifs. Ce résultat fait mentir la conjecture de Harris selon laquelle la prise en compte de l'énoncé complet est la condition garantissant que l'inférence porte sur un conditionnement morphologique plutôt que phonologique ou syllabique notamment (Harris 1955, pp.59-60). Notre premier objectif dans ce mémoire sera donc de dériver des propositions initiales de Harris un modèle markovien - probabiliste en particulier - de la segmentation, et de clarifier le statut relatif des mesures de diversité pouvant jouer le rôle de critère dans ce contexte. C'est ce modèle que nous chercherons à étendre au cas plus général d'une approche basée sur la similarité entre la distribution des successeurs d'un

9. Nous appelons *variété conditionnelle* l'équivalent du nombre de successeurs sous les hypothèses markoviennes (voir alinéa 2.2.2.1).

10. sans que les auteurs semblent s'en apercevoir réellement

11. ce n'est pas le cas de la variété conditionnelle

contexte et celle des phonèmes en début d'énoncé[12].

Par un détour trivial de la pensée, maximiser la similarité revient à minimiser la différence. Dans le contexte markovien esquissé ci-dessus, la Théorie de l'Information justifie qu'on évalue la différence entre deux distributions[13] par leur *divergence* (*de Kullback-Leibler*) ou entropie *relative*, dont la définition est beaucoup plus proche de celle de l'entropie (de Shannon) que ne le laisse présumer l'écart conceptuel qui sépare la notion de (dis-)similarité de celle d'incertitude. Outre le fait de n'être guère plus compliquée à comprendre et à utiliser que l'entropie, la divergence présente plusieurs propriétés désirables pour notre propos, dont la plus fondamentale est qu'elle s'annule ssi les deux distributions sont identiques, en l'occurrence si la distribution conditionnelle en un point concorde exactement avec la distribution initiale (voir note 12, page précédente). De plus, elle ne tient pas compte, dans son évaluation globale de la différence entre les distributions, des cas où la probabilité conditionnelle d'un symbole est nulle alors que son homologue initiale ne l'est pas, ce qui traduit l'hypothèse *tolérante* que la nullité de la probabilité conditionnelle peut être le fait d'un défaut de la procédure d'échantillonnage; à l'inverse, la divergence est infinie si l'une au moins des probabilités conditionnelles est non-nulle tandis que la probabilité initiale correspondante est égale à zéro - c'est-à-dire que l'apparition dans un contexte donné d'un symbole interdit en position initiale aboutit à la réfutation systématique d'un modèle prédisant l'insertion d'une frontière dans ce contexte.

A ce point de l'exposé, un lecteur sensible à ce genre d'arguments aura probablement déjà formé l'intuition que le second objectif de ce mémoire sera de justifier et formaliser dans une perspective markovienne la conception d'une segmentation basée sur la divergence comme généralisation de la segmentation entropique[14]. Si nous y parvenons, nous pourrons nous tourner vers notre dernière et plus ambitieuse interrogation: comment estimer la distribution initiale en l'absence de segmentation dans le corpus? Dans le cas de langues telles que l'anglais, on y répond généralement en faisant (implicitement) l'hypothèse que la distribution en question est uniforme, d'où l'utilisation de l'entropie conditionnelle - qui se trouve être inversement proportionnelle à la divergence entre la distribution conditionnelle et la distribution uniforme correspondante. Nous

12. Dans la suite, nous référerons à la première sous le titre de distribution *conditionnelle*, et à la seconde sous celui de distribution *initiale* (ainsi que pour les probabilités correspondantes); il importe de comprendre que la première varie en fonction de la position considérée (pour l'insertion d'une frontière), tandis que la seconde est fixée pour l'ensemble de la séquence segmentée.

13. plus précisément, entre deux fonctions de probabilité sur la même collection d'événements

14. telle que décrite par Hutchens & Alder (1998) notamment

proposons de substituer à cette hypothèse celle, plus générale, que la distribution initiale peut être approximée par la distribution (que nous qualifierons d'*inconditionnelle* à la seule fin d'éviter des confusions) des phonèmes de la langue; pour aller plus loin, nous avançons qu'en l'absence d'autres informations sur les données, la distribution inconditionnelle est notre meilleure approximation de la distribution initiale. Cette nouvelle hypothèse est justifiée par le fait que la divergence entre les distributions conditionnelle et inconditionnelle correspond à l'*information mutuelle moyenne* associée au contexte considéré, qui s'interprète comme une mesure de la contrainte qu'il impose sur sa distribution conditionnelle. En d'autres termes, utiliser la distribution inconditionnelle comme approximation de la distribution initiale dans le calcul de la divergence (et la divergence comme critère pour l'insertion de frontières) revient à segmenter la chaîne parlée aux points où la distribution conditionnelle est (autant que possible) indépendante du contexte.

Pour mettre à l'épreuve des données le raisonnement résumé ici et repris en détail dans les partie 2 et 3, nous avons effectué une expérimentation informatique portant sur deux corpus orthographiques simplifiés (latin et anglais) de plusieurs centaines de milliers de symboles. Nous avons cherché à évaluer l'adéquation, pour la segmentation en *mots*, des quatre critères mentionnés précédemment: variété et entropie conditionnelle, information mutuelle moyenne et divergence entre distribution conditionnelle et initiale, cette dernière étant approximée ici par la véritable distribution en début de mot. Les résultats, présentés dans la partie 3, vont dans le sens de nos hypothèses initiales; en particulier, pour la tâche et les corpus considérés, on constate que:

- la divergence donne toujours de meilleurs résultats que l'entropie;
- l'information mutuelle donne toujours de meilleurs résultats que l'entropie en latin;
- l'information mutuelle ne fait jamais beaucoup moins bien que l'entropie en anglais;

Nous en concluons d'une part que, si l'on connaît la *vraie* distribution initiale, il est possible d'améliorer considérablement les résultats du système; d'autre part, qu'il y a de bonnes raisons d'adopter le critère du minimum de dépendance si l'on n'a pas de raison de penser que le critère du maximum d'incertitude est plus approprié pour les données considérées.

Au moment de dresser le bilan du chemin parcouru, nous chercherons à

mettre en avant les paramètres du modèle qui restent à examiner, les extensions envisageables et les problèmes en suspens. En particulier, nous tenterons d'esquisser les contours d'un algorithme de segmentation plus sophistiqué, exploitant la dichotomie entre distribution conditionnelle et initiale dans la perspective d'un ajustement itératif des paramètres de la seconde.

Chapitre 1

Introduction

Dans son article «From Phoneme to Morpheme» (Harris 1955), Harris décrit une procédure permettant d'inférer la position des frontières morphologiques dans un énoncé sur la base de l'observation de la distribution des séquences de phonèmes dans la langue considérée. Depuis lors, la recherche sur le sujet semble ne s'être jamais tarie, et de nombreuses variantes sont régulièrement proposées - le plus souvent de l'ordre d'un paramétrage subtil du modèle initial; au-delà des spécificités algorithmiques, la majorité de ces méthodes ne remettent pas en question l'hypothèse fréquemment attribuée à Harris [1] selon laquelle l'*incertitude* sur l'enchaînement des phonèmes est liée à la position des frontières d'unités morphologiques. C'est dans ce paradigme que vise à s'insérer le présent travail; en particulier, nous chercherons à justifier l'utilisation d'un critère plus général que l'incertitude, à savoir l'*indépendance* des unités successives, sur la base de considérations apparaissant dans l'article séminal de 1955 mais dont personne ne semble avoir réellement tenu compte par la suite.

Pour les personnes auxquelles l'idée de compter les unités (ou pire, les transitions d'unités) d'un corpus est venue assez naturellement, les lignes qui précèdent devraient suffire à poser un contexte assez précis pour la discussion. En proportion du genre humain, cette catégorie est plutôt minoritaire [2]; elle correspond à l'intersection de deux ensembles (un peu) plus vastes: d'une part les linguistes, dont il est vraisemblable qu'ils conçoivent aisément l'importance et la complexité de la détermination des unités morphologiques, et de l'autre les

1. Cette attribution est incorrecte, ainsi que nous l'avons indiqué dans l'avant-propos et que nous chercherons à le mettre en évidence par la suite.

2. encore que le syndrome tende à se répandre de façon inquiétante depuis une vingtaine d'année

statisticiens, pour lesquels indépendance et incertitude font l'objet d'une réflexion de chaque instant. Les personnes appartenant exclusivement à l'une *ou* l'autre des corporations devront peut-être, avouons-le, dépasser certaines réticences pour traverser le texte sans douleur.

Elles s'en consoleront facilement en considérant l'embarras probable du néophyte, dont la perplexité face à ce genre de considération est souvent exprimée par une question simple et fondamentale: *à quoi ça sert ?* Implicitement, son intuition est que la segmentation morphologique des énoncés est transparente et univoque, et la présente réflexion parfaitement redondante. C'est le propos du premier et plus important paragraphe de cette introduction que de dissiper ce sentiment en donnant quelques contre-exemples empruntés à divers domaines de la recherche en linguistique théorique et appliquée. Le paragraphe suivant sera consacré à une première discussion de l'ensemble de contraintes que nous imposons à un modèle de segmentation morphologique, et qui imposent à leur tour des limites absolues à ses performances. Alors seulement nous tenterons d'expliciter - de façon informelle d'abord, la nature du développement que nous proposons et les arguments théoriques en faveur de son adoption.

1.1 Indétermination formelle des frontières morphologiques

Dans une première approximation, la segmentation morphologique peut être définie comme l'opération consistant à déterminer la position des frontières des (signifiants de[3]) *signes morphologiques* au sein d'une séquence de phonèmes d'une langue[4]. Le «Cours de Morphologie Générale» (CMG) de Mel'čuk définit le *signe* linguistique comme l'association d'un *signifié*, un *signifiant* et un *syntactique* (ou combinatoire) constants (Mel'čuk 1997, p.15). Les signes *morphologiques* sont ceux parmi les signes linguistiques qui ne peuvent être décrits en termes des relations syntaxiques existant entre leurs constituants; rien n'exclut en revanche qu'ils soient décomposables en d'autres signes morphologiques et ainsi de suite récursivement, jusqu'à ce que la condition d'*élémentarité* du signe soit satisfaite, c'est-à-dire qu'on ne puisse pas le décomposer plus avant dans la langue considérée sans nuire à son caractère *significatif*. Si l'on restreint son

3. La notion de *frontière* n'a aucun sens si c'est de signe que l'on parle; elle ne peut s'appliquer qu'au segment phonologique qui forme son signifiant.

4. Par extension, nous parlerons quelquefois de *segmentation* d'un corpus donné pour désigner le *résultat* de l'application de la segmentation (comme *opération*) à ce corpus.

attention aux signes *segmentaux*[5], on peut distinguer en particulier les *morphes* ou signes morphologiques *élémentaires* et les *mots-forme*[6]:

> *Grosso modo*, on peut dire qu'un mot-forme typique est construit de morphes et qu'on peut donc, en règle générale, le «décomposer» en morphes constituants (pour ainsi dire, le «découper» en tronçons morphiques) ou le «construire» par simple juxtaposition des morphes. Un mot-forme, ou chaîne de morphes, constitue en effet le cas le plus fréquent de signe linguistique$_1$ non élémentaire, de sorte que nous pouvons postuler l'équation suivante, qui est valide dans la plupart des cas:
>
> mot-forme = morphe$_1$ + morphe$_2$ + ... + morphe$_n$
>
> (Mel'čuk 1997, p.351, souligné par l'auteur[7])

En français, par exemple, la séquence /ilpøʁəpartir/ *il peut repartir* peut être divisée en trois sous-séquences, /il/, /pø/ et /ʁəpartir/, qui sont les signifiants de trois mots (voir note 20). Le dernier peut à son tour être segmenté en trois signifiants de morphes: /rə-part-ir/, que l'on retrouve dans une multitude de cas avec la même signification, le même signifiant et les mêmes contraintes et latitudes combinatoires: /ʁə-vjɛ̃/ *reviens/-t*, /paʁt-ɔ̃/ partons, /sɔʁt-iʁ/ sortir et ainsi de suite. Pour être précis, c'est *en vertu de* ce que ces éléments manifestent des propriétés permanentes dans des contextes variés qu'on peut les identifier comme des unités linguistiques et en particulier morphologiques du français.

La segmentation morphologique serait un problème trivial pour la description des langues si les frontières morphologiques étaient systématiquement marquées formellement dans la chaîne parlée. Or cette contrainte est régulièrement violée dans les langues naturelles, où l'on observe que les pauses sont plus généralement situées entre des groupes de mots qu'entre les mots eux-mêmes - naturellement, les frontières de morphes qui ne sont pas des frontières de mots ne sont pour ainsi dire jamais marquées de cette façon[8].

---

5. Un signe est dit *segmental* si son signifiant est une séquence de phonèmes; c'est généralement le cas dans les langues naturelles (voir par exemple Mel'čuk 1997, p.11), mais il existe une quantité de signes dont le signifiant est un trait prosodique ou une modification du signifiant d'un autre signe, par exemple.

6. L'utilisation du terme *mot-forme*, si elle correspond à une distinction justifiée dans le CMG, n'a pas de raison d'être pour notre propos; nous lui préférons dans le reste de ce travail le *mot* traditionnel.

7. La notation indicée «linguistique$_1$» est caractéristique du style de Mel'čuk; elle ne nécessite pas d'éclaircissement particulier dans notre contexte.

8. Ici comme dans l'ensemble de ce mémoire, nous restreignons notre propos à une morphologie purement segmentale (voir paragraphe suivant), ce qui exclut de nombreux phénomènes prosodiques ou accentuels dont on sait pourtant qu'ils sont souvent de bons indices pour la segmentation (voir par exemple Brent & Cartwright 1995).

Cette indétermination formelle s'étend au plan du langage écrit. L'écriture des langues indo-européennes implique généralement l'usage de symboles marquant explicitement les limites de mots; ainsi, dans l'énoncé *j'ai congé aujourd'hui*, les espaces jouent clairement le rôle de séparateurs de mots. Mais la fonction de l'apostrophe est plus ambiguë, puisqu'il est séparateur dans le contexte *j'avais* et non dans *aujourd'hui*. Bien entendu, ces langues n'ont généralement pas de symbole explicitant les frontières de morphes à l'intérieur d'un mot. A l'extrême, la graphie d'autres langues comme le chinois ou le japonais ne comporte aucun séparateur explicite; dans ce dernier cas, l'ambiguïté est plus élevée encore qu'elle ne l'est à l'oral[9].

La segmentation morphologique fait également l'objet d'un débat central dans le domaine des théories de l'acquisition du langage. En somme, les difficultés que rencontre le linguiste dans la pratique de la segmentation sont comparables à celle d'un enfant apprenant sa première langue. Les deux situations diffèrent par la quantité et la complexité des informations méta-linguistiques à disposition de chaque partie, ainsi que par la nature plus ou moins pédagogique du *corpus*[10]: le langage au moyen duquel les parents communiquent avec leur enfant est en quelque sorte *optimisé* pour l'acquisition, ce qui se traduit d'une part par l'utilisation de structures linguistiques simplifiées, dont les contrastes sont souvent «artificiellement» accentués par des traits prosodiques, et d'autre part par le degré élevé de prédictibilité de son déroulement[11]. Ces deux aspects de l'input parental correspondent à deux types d'indices formels dont on présume que l'enfant se sert pour inférer les frontières morphologiques: indices suprasegmentaux et indices liés à la distribution des phonèmes au fil de la séquence.

Les deux approches ont été explorées systématiquement, en particulier au cours de la dernière décennie, et il semble qu'elles soient partiellement indépendantes, en ce sens que l'ensemble des frontières morphologiques correctement prédites par l'une ne se confond pas strictement avec celui de l'autre[12]. De façon générale, on constate que les indices suprasegmentaux utilisés pour marquer

9. On peut citer par ailleurs le cas de graphies antiques où les symboles se succèdent sans séparateurs, ou encore la pratique usuelle en cryptographie consistant à supprimer les espaces du texte à coder, leur fréquence élevée les rendant trop aisément détectables; notons que, dans ces deux exemples, les symboles ne correspondent pas à des morphes, contrairement aux idéogrammes.

10. Dans le contexte des théories de l'acquisition, on parle traditionnellement d'*input* (*parental*).

11. par opposition au langage littéraire, par exemple, où la diversification des structures linguistiques constitue un objectif stylistique explicite

12. ce qui n'implique pas nécessairement leur disjonction, par ailleurs

la segmentation dépendent spécifiquement des langues considérées, tandis que le fait que la position des frontières soit au moins partiellement conditionnée par la distribution des phonèmes en divers points de la chaîne est une propriété très générale des langues naturelles.

L'avantage certain des critères suprasegmentaux réside dans le rapport entre la régularité de leurs prédictions et la concision de leurs représentations. Ainsi, dans de nombreuses langues, on observe que quelques règles accentuelles[13] simples rendent compte de la segmentation morphologique bien mieux que ne le feraient les nombreux paramètres d'un modèle statistique voué à la même tâche. En revanche, leur statut *spécifique* (pertinent uniquement dans le cadre d'une langue donnée) justifie que l'on conçoive la capacité d'effectuer la segmentation sur une base distributionnelle comme un aspect plus fondamental du langage humain, en dépit de sa complexité apparemment plus élevée; les résultats de plusieurs expérimentations convergent dans cette direction:

> [...] adults are able to discover word units rapidly even in a system as impoverished as an unsegmented artificial language. This result is rather remarkable, given that such distributional cues have generally been considered to be too complex for human learners to use in language learning. (Saffran, Newport & Aslin 1996, p.618)

La perspective adoptée dans ce mémoire est plus proche des ambitions descriptives évoquées précédemment que de celles d'une étude de l'acquisition des unités morphologiques. Toujours est-il que cette discipline fournit, comme on le voit, un argument pour la validité de l'approche purement distributionnelle qui sera appliquée plus loin. Cela ne signifie naturellement pas qu'on ait intérêt à négliger les indices autres que distributionnels pour la segmentation, mais qu'on peut s'attendre à des résultats intéressants même en leur absence.

Enfin, le domaine du traitement automatique du langage est concerné de diverses façons par la définition d'algorithmes de segmentation morphologique. On peut notamment citer des applications dans le cadre de la reconnaissance de la parole, de la correction et de l'acquisition lexicale. Au vu du problème de l'indétermination des frontières morphologiques à l'oral, l'importance de la segmentation pour la reconnaissance de la parole est évidente et nous pouvons faire l'économie de cette discussion au profit des deux autres.

Le problème de la *correction lexicale* peut être résumé comme suit: étant donné un lexique $L := \{w_1, \ldots, w_l\}$ et un mot inconnu $w \notin L$, comment définir une fonction $f(w_i)$ associant à tout mot du lexique $w_i \in L$ une valeur indiquant son adéquation comme forme correcte de la forme supposément erronée w. Bien

13. typiquement, mais aussi tonales, relatives à la durée des phonèmes, etc.

entendu, ce modèle n'est pas entièrement spécifié tant que la notion d'*adéquation* reste à expliciter, mais là n'est pas notre propos. Dans ce contexte, la segmentation fournit un moyen d'éviter que la faute typographique fréquente consistant à omettre l'espace entre deux mots entraîne des propositions de correction aberrantes, comme par exemple *maison* pour *mais=il* (*mais il*).

La désignation d'*acquisition lexicale* (*non-supervisée*) est généralement réservée à l'opération consistant à identifier, de façon aussi automatique que possible [14], les unités morphologiques qui composent un corpus textuel, typiquement en vue de constituer un lexique (le plus souvent de mots). Notons que, même pour des systèmes graphiques utilisant des séparateurs explicites, le problème est loin d'être trivial. Nous avons déjà mentionné que le statut de séparateur de l'apostrophe est «irrégulier» (en français notamment), et c'est aussi le cas du tiret (*peut-être - peut-il*), du point et de la virgule (particulièrement en ce qui concerne les sigles et abréviations, ainsi que les notations numériques); l'espace peut également ne pas fonctionner comme séparateur dans certains cas (*parce que*, *pomme de terre*, etc.) [15]. Dans l'ensemble, le statut des symboles non alphabétiques est plus complexe qu'il n'y paraît à première vue, et cette ambiguïté entraîne des répercussions sur toutes les méthodes de traitement automatique de l'écrit qui impliquent la segmentation en mots [16] des corpus considérés.

Au terme de cette première approche, nous espérons avoir montré que la segmentation morphologique est à la fois une tâche complexe et une condition basique de nombreux développements linguistiques. Elle implique une telle quantité d'informations sur la langue qu'un modèle formel ne peut en donner au mieux qu'une bonne approximation. Dans le paragraphe suivant, nous chercherons à énoncer plusieurs contraintes sur la forme du modèle, dont on peut s'attendre à ce qu'elles induisent un certain nombre d'erreurs, mais qui devraient permettre de ramener la dimensionnalité du problème de la segmentation à des proportions plus ou moins contrôlables.

1.2 Contraintes générales pour un algorithme de segmentation

Préalablement à une description plus formelle du modèle de la segmentation que nous étudions dans cette recherche, nous souhaitons discuter ici le jeu de

14. c'est-à-dire dire impliquant aussi peu d'informations méta-linguistiques que possible

15. Voir Manning & Schütze (1999, pp.124-31) pour une discussion plus spécifique des problèmes en question.

16. souvent appelée *tokenization* dans la littérature scientifique anglophone

simplifications conceptuelles sur lequel reposent les développements théoriques subséquents et qui caractérise le signifiant des unités morphologiques comme un objet segmental, continu, linéairement séparable et manifeste - autant de conditions qui sont fréquemment infirmées dans les langues naturelles, mais dont il est raisonnable d'avancer qu'elles correspondent à la forme de base du morphe. Nous chercherons en particulier à mettre en évidence la justification de ces contraintes tout en relevant les limites absolues qu'elles impliquent pour le traitement des langues naturelles.

Nous avons déjà mentionné que la portée de nos considérations est restreinte au caractère purement *segmental* du signifiant morphologique (voir notes 5 et 8, p. 3 et 3). Il est clair qu'on peut trouver une foule de contre-exemples dans les langues du monde; sans donner dans l'exotisme ou la rareté linguistique, le signe exprimant la modalité interrogative est fréquemment exprimé en français oral (notamment) par l'intonation ascendante appliquée aux dernières syllabes de l'énoncé. Ce genre de signe - et d'autres encore pour les mêmes raisons, échappent complètement à l'analyse que nous proposons ici. Notons toutefois avec Mel'čuk que:

> Le MORPHE est le signe morphologique de loin le plus important dans les langues du monde. Il est universel en ce sens [...] qu'aucune langue ne se passe de morphes. Quantitativement, les morphes constituent quelque chose comme 99% du stock des signes morphologiques dans toute langue; c'est, de façon universelle, le pain quotidien de la communication linguistique$_1$.
> (Mel'čuk 1997, p.11, souligné par l'auteur)

On voit ainsi que la prééminence du niveau segmental est un caractère *général* des langues naturelles, et ce constat relativise (sans l'annuler) le défaut de généralité de l'approche envisagée.

Cette première contrainte exclut de la discussion un ensemble important de procédés morphologiques; elle est une condition nécessaire, mais non suffisante, à la formalisation que nous envisagerons plus loin. Nous devons également insister ici sur les aspects continu et linéairement séparable que nous postulons pour le signifiant des unités morphologiques. Par *continuité*, nous entendons que les symboles constituant le signifiant d'une unité doivent être strictement adjacents. A nouveau, les signes morphologiques ne sont pas rares qui font mentir cette hypothèse. L'interruption d'un signifiant par un autre constitue d'ailleurs le fondement d'une classification répandue des affixes[17]: on distingue ainsi les *confixes*

17. Les morphes d'une langue sont traditionnellement répartis en deux catégories basiques: les *racines* forment la classe la plus vaste et la plus variable, et sont caractérisées en premier

(*préfixes*, *suffixes* et *interfixes*), qui n'interrompent pas le signifiant des éléments auxquels ils s'appliquent, et les autres types de morphes affixaux, qui ne vérifient pas cette propriété: *circonfixes*, *infixes* et *transfixes* (Mel'čuk 1997, pp.147-87). Par exemple, le participe passé allemand est très régulièrement formé par circonfixation: *ge-frag-t*, *ge-leg-t*, etc. De même, l'arabe est connu pour son usage intensif des transfixes flexionnels[18]. En pareil cas, nous n'attendons rien de plus d'un algorithme de segmentation qu'une détection correcte des *frontières* morphologiques des signifiants considérés. Ici encore, on peut justifier partiellement ce sacrifice en relevant que les confixes sont considérablement plus fréquents que les autres classes dans les langues naturelles (Mel'čuk 1997, p.149)

Le caractère *linéairement séparable* des unités désigne l'idéalisation selon laquelle chaque élément de la chaîne phonologique peut être attribué à un signifiant et un seul. Ce postulat est clairement invalidé par l'observation des phénomènes d'assimilation des frontières à l'oral, dont l'anglais américain fournit des exemples célèbres: /wonšʌ/ *wontcha* pour /wontjʌ/ *won't ya,* /donšʌ/ *dontcha* pour /dontjʌ/ *don't ya*, etc[19]. Dans ces cas, les frontières disparaissent, pour ainsi dire, à l'intérieur du phonème /š/ substitué à la paire /tj/. On rencontre le même genre de problème lorsque plusieurs signes sont réalisés simultanément par un signifiant unique: ainsi, en français, on serait bien emprunté de segmenter la forme /o/ *au* qui se substitue obligatoirement à la séquence «régulière» /alə/ *à le*, ou le suffixe verbal /õ/ *-ons* qui indique à la fois la personne et le nombre. Dans notre perspective, la segmentation consiste en l'insertion de frontières entre les constituants d'une paire de symboles, ce qui implique que les segments considérés /wonšʌ/, /donšʌ/, /o/ et /õ/ sont conçus comme des signifiants uniques.

En définissant le morphe comme unité linguistique *manifeste*, nous visons à exclure du champ d'investigation les signes morphologiques qui, tout en participant au sens d'un énoncé, n'y figurent pas *formellement* - si ce n'est par

lieu par l'importance de leur combinatoire *interlexémique*, c'est-à-dire spécifiant les modalités de leurs relations avec d'autres mots dans un énoncé; à l'inverse, les *affixes* forment une classe plus réduite et plus figée, et leur syntactique est fortement orienté vers la combinatoire *intralexémique*, donc la cooccurrence des morphes dans les limites du mot (voir notamment Mel'čuk 1997, pp.62-75).

18. La flexion dans cette langue est réellement déroutante pour qui a construit sa compétence linguistique en baignant dans le suffixe indo-européen; par exemple, les racines consonantiques /bjt/ (maison) et /rsm/ (dessin) se combinent avec les transfixes vocaliques /-a-∅-/ et /-u-ū-/ pour distinguer le singulier (/bajt/, /rasm/) du pluriel (/bujūt/, /rusūm/) (Mel'čuk 1997, p.175).

19. Le symbole /š/ utilisé ici représente l'affriquée sourde anglaise dont la réalisation *phonétique* est notée [tʃ].

leur absence relativement à d'autres membres d'un paradigme fini. On parle généralement de signifiant *zéro* pour désigner cette configuration plus courante qu'on pourrait le penser, en particulier dans les langues où les significations grammaticales (genre, nombre, temps, mode, etc.) sont typiquement exprimées par affixation. Ainsi, en français, le morphe exprimant le temps verbal dans la séquence /tyʃãt/ *tu chantes* a un signifiant zéro, comme l'indique la comparaison avec l'imparfait /tyʃãtɛ/ *tu chantais* [20]. Cela n'affecte pas la position des frontières, mais le fait de renoncer à l'information de la présence d'un signe zéro, à l'instar des contraintes déjà mentionnées, contribue à différencier l'opération de segmentation telle qu'elle est définie ici d'une véritable analyse morphologique [21].

Pour rendre justice à la théorie distributionnaliste, nous devons signaler que toutes ces questions et d'autres encore sont discutées de façon approfondie dans les articles du recueil «Papers in Structural and Transformational Linguistics» (Harris 1970, voir en particulier la première partie, pp.3-157); ce n'est que par manque de temps [22] que nous renonçons à les prendre en considération ici. Nous pensons toutefois avoir assez insisté sur la fréquence, dans les langues naturelles, des faits morphologiques subsistant, ceux dont peut rendre compte le modèle basique décrit dans ce mémoire. Ainsi, s'il est évident que le modèle en question est *irréaliste*, il ne paraît pas audacieux de nuancer la qualification en précisant qu'il correspond à une conception *élémentaire* - et non uniquement *opérationnelle* - de la segmentation.

1.3 Incertitude, divergence, indépendance

Pour conclure cette introduction, et en anticipant quelque peu sur la suite de l'exposé, nous tenterons de donner une expression informelle du développement spécifique que nous suggérons d'apporter au modèle initial de la segmentation tel qu'il est défini par Harris (1955) [23]. Nous avons mentionné plus haut le statut fondamental de la notion d'*incertitude* dans ce modèle; en fait, il serait plus exact

20. Cette comparaison ne fait office de preuve que si l'on admet la conception, défendue notamment par Mel'čuk (1998, p.43-9), selon laquelle les signes linguistiques en général et les morphes en particulier ont par définition un caractère strictement *additif*.

21. Il convient d'ajouter à ce propos que l'identification des relations d'*allomorphie* (comme celle qui existe en anglais entre les racines signifiant «épouse»: /wa^{i}f/ *wife* et /wa^{i}v-z/ *wives*, où la notation /a^{i}/ représente le phonème anglais transcrit phonétiquement par [ai]) et d'*homonymie* (c'est le cas, en français, du signifiant /fɛr/, qui correspond à deux signes distingués à l'écrit: *faire* et *fer*) échappent complètement au formalisme décrit dans ces pages.

22. et, nous l'espérons, provisoirement

23. ou plutôt, tel qu'il sera reformulé au paragraphe 2.2

de dire que la notion de *diversité* est à la base de cette approche. Dans les grandes lignes, le principe consiste à inférer la présence d'une frontière morphologique aux points de la chaîne parlée ou la structure de la langue, telle qu'elle nous est connue par le biais d'un modèle, impose (localement) le moins de contraintes sur la distribution des phonèmes successifs.

Pour illustrer cette idée, considérons le cas d'un automate[24] aléatoire conçu pour émettre, phonème par phonème, une séquence infinie de mots français, et supposons qu'on l'ait doté d'un lexique à cette fin. A chaque itération, l'automate est amené à sélectionner un symbole parmi ceux qu'autorise le lexique à cette position; par exemple, s'il a émis les phonèmes /t/, /a/ et /b/ en début de mot[25], il doit choisir entre six successeurs possibles: les voyelles /y/, /ɛ/, /ɔ/, /a/ et /u/, et la consonne liquide /l/ (on observe notamment les formes /tabyle/ *tabuler*/*-é*, /tabɛs/ *tabès*, /tabɔr/ *tabor*, /taba/ *tabac*, /tabu/ *tabou* et /tabl/ *table*[26]). S'il retient le /a/, il a la possibilité de terminer le mot courant pour en commencer un nouveau, mais il peut aussi poursuivre avec l'une des consonnes suivantes: /t/, /s/, /ʒ/ et /r/ (d'après /tabatiɛr/ *tabatière*, /tabas/ *tabasse*, /tabaʒi/ *tabagie* et /tabar/ *tabar*). S'il sélectionne alors le /t/, les phonèmes subséquents sont entièrement déterminés par la structure de la langue telle qu'elle est définie dans le lexique: il ne subsiste aucune autre option que de terminer la forme /tabatiɛr/, avant de pouvoir à nouveau effectuer un choix. Comme on le voit, le nombre des successeurs possibles en un point de la chaîne n'est pas constant. Il est généralement élevé aux frontières d'unités morphologiques, et tend à diminuer tant que l'on progresse à l'intérieur d'un signifiant; c'est ce qui justifie la proposition de Harris[27].

La notion d'incertitude n'apparaît pas en tant que telle dans sa formulation de sa méthode, mais on comprend intuitivement qu'elle croît avec le nombre de successeurs, et l'on verra dans l'alinéa 2.2.3.2 (et l'annexe A.2) qu'elle présente d'autres propriétés intéressantes. Il n'est pas surprenant, dès lors, qu'on ait tenté de développer le modèle initial dans ce sens. Ce qu'il l'est plus, c'est qu'on n'ait jamais vraiment tenu compte, ce faisant, d'une remarque fondamentale de Harris: il indique en effet que le principe fondateur de la méthode

24. au sens usuel ou technique, à la convenance du lecteur

25. et si son lexique (phonologique) est basé sur les indications de l'édition 1989 du petit Larousse illustré

26. Il importe de noter que la contrainte structurelle qui aboutit à cette énumération est d'ordre *lexical*, plutôt que seulement *phonotactique*, puisqu'on atteste les enchaînements /bi/, /br/, /bd/, etc. à d'autres positions à l'intérieur de mots (par exemple /abi/ *habits*, /abri/ *abri*, /abdomɛn/ *abdomen*).

27. Nous reviendrons plus loin sur certaines des extensions envisagées par Harris dès son article de 1955; elles ne sont pas directement pertinentes pour la présente discussion.

n'est pas la maximisation de la diversité ou de l'incertitude sur la distribution conditionnelle[28] en un point, mais la maximisation de la similarité entre cette distribution et celle (dite initiale) attendue en fin d'énoncé (*utterance*) dans la langue considérée - donc dans le seul contexte où l'on est certain de traverser une frontière morphologique, en l'absence d'autre information *a priori*. Dans le cas de langues telles que l'anglais, c'est-à-dire où la similarité en question est bien approximée par la seule considération du nombre de phonèmes différents pouvant apparaître, nous pouvons postuler que la distribution initiale est uniforme[29] et utiliser au lieu d'une mesure de similarité entre les deux distributions une mesure de diversité ou plus spécifiquement d'incertitude sur la distribution conditionnelle uniquement.

La première hypothèse que nous posons dans ce mémoire (et qui présente tous les caractères d'un *retour aux sources*) est que, dans les langues où la distribution initiale s'écarte de l'uniformité, il devrait être possible d'obtenir une meilleure segmentation en se basant sur un indice de (dis-)similarité[30] si nous disposons du *vrai* modèle de la distribution initiale; comme on l'a dit plus haut, la seule surprise à propos de cette idée est qu'elle n'ait pas été explicitée jusqu'ici (à notre connaissance) en dehors des travaux de Harris. Elle présente l'intérêt de différencier explicitement la contribution des deux modèles au processus de segmentation, et du même coup les modalités de l'acquisition de leurs paramètres; comme on sait, depuis l'expérience conduite par Harris (1967), qu'il est possible d'estimer les paramètres de la distribution conditionnelle à partir du corpus que l'on s'apprête à segmenter, il paraît légitime de se concentrer sur le problème plus spécifique de l'acquisition, à partir du même corpus «brut», d'un *bon* modèle de la distribution initiale - en ce sens qu'il s'approche autant que possible de la *vraie*.

C'est sur ce point que porte notre seconde hypothèse: nous avançons qu'en l'absence de précision complémentaire, la distribution des symboles *hors contexte*, que nous nous permettons de qualifier ici de distribution *inconditionnelle*, constitue une approximation justifiable de la distribution initiale - la meilleure, à vrai dire, qu'on puisse tirer d'un corpus non segmenté. Nous verrons dans l'alinéa 2.1.2.4 (et l'annexe A.2) que, dans les termes de la Théorie de l'Information, la *divergence* (voir note 44) entre la distribution conditionnelle des successeurs

28. Voir note 12.

29. ou du moins que tous les symboles de l'alphabet y figurent, si l'on parle de diversité plutôt que d'incertitude

30. En l'occurrence, nous verrons que l'indicateur de prédilection dans ce contexte est la *divergence* (*de Kullback-Leibler*) ou entropie *relative* entre deux distributions (voir paragraphe 3.1 et annexe A.2).

d'un contexte donné et celle (inconditionnelle) des phonèmes de la langue est appelée *information mutuelle moyenne* et mesure traditionnellement la *dépendance* entre le contexte considéré et ses successeurs possibles. Ainsi, approximer la distribution initiale par la distribution inconditionnelle revient à segmenter la chaîne par disjonction des unités successives les moins dépendantes.

La suite de l'exposé sera structurée en fonction des hypothèses formulées dans ce paragraphe. La partie 2 sera consacrée à l'exposition du modèle initial de Harris et à la dérivation d'un modèle comparable exprimé en terme de chaînes de Markov d'ordre k. Nous examinerons en particulier la possibilité d'inclure dans le calcul de notre mesure de diversité la fréquence des transitions considérées, et tenterons d'intégrer au formalisme la proposition de Harris de tenir compte des transitions *vers l'arrière* (*backward*), que nous appelons *précédences*. Dans la partie 3, nous complexifierons le modèle en introduisant successivement ses versions basées sur la divergence et sur l'information mutuelle. Nous donnerons alors les résultats d'une comparaison systématique des indicateurs exposés dans le cadre d'une application de segmentation en mots de corpus orthographiques simplifiés anglais et latin (partie 4), qui nous permettront de jauger la validité de nos hypothèses avant de conclure par une discussion forces et faiblesses de l'approche, et des développements suggérés par nos résultats.

Chapitre 2

Du nombre de successeurs à l'entropie conditionnelle

Après une longue introduction aux enjeux de la segmentation morphologique, nous pouvons aborder la formalisation du modèle que nous mettrons plus loin à l'épreuve des données. La présentation sera organisée en trois étapes.

Dans un premier temps nous définirons un algorithme générique de segmentation par *seuillage*, que nous illustrerons par la méthode du *nombre de successeurs* telle qu'elle est définie dans Harris (1955); nous verrons en particulier que cette classe de procédures est caractérisée par la conception de la segmentation comme un problème de détection de signal, c'est à dire de réaction à un *stimulus* au-delà d'une *perceptibilité* donnée.

Puis nous formulerons plusieurs hypothèses quant au conditionnement du *critère* (ou *variable de décision*, dont le nombre de successeurs est un exemple) et du *seuil* au-delà duquel le critère est tenu pour significativement élevé. Ces considérations nous permettront de donner une expression du modèle en terme de *chaîne de Markov* d'ordre k.

Nous introduirons alors successivement plusieurs critères utilisés (ou du moins utilisables) dans ce contexte, en nous référant aux éléments formels donnés dans l'annexe A. Nous présenterons ainsi l'équivalent markovien du nombre de successeurs, la *variété conditionnelle*, et tenterons de justifier le pas supplémentaire consistant à prendre en compte les fréquences pour obtenir une mesure d'incertitude, l'*entropie* conditionnelle. Enfin, nous proposerons une façon de combiner les informations portant sur le déroulement *vers l'avant* de la séquence et celles portant sur le déroulement *inverse*, de la fin de la séquence vers le début.

2.1 Segmentation par seuillage: la méthode du nombre de successeurs

La segmentation par seuillage est conçue comme un problème de détection de signal[1]. Elle s'oppose en particulier à deux autres classes majeures d'algorithmes de segmentation (non supervisée):

- les techniques *ascendantes* (*bottom-up*) procédant par *agrégation* des co-occurrences fréquentes, dont les implémentations les plus fameuses sont celles de Wolff (1977) et Nevill-Manning (1996); le principe fondateur de ces approches est de considérer l'alphabet comme un lexique de base, et de recoder des paires d'entrées du lexique comme des entrées autonomes dès lors qu'elles excèdent une fréquence donnée;

- les techniques *descendantes* (*top-down*) consistant à sélectionner une segmentation parmi toutes celles possibles pour un corpus, généralement sur la base d'un critère de *compression* (voir par exemple Brent & Cartwright (1996), de Marcken (1996), Kit & Wilks (1999)) ou de *maximum de vraisemblance* (voir notamment Olivier (1968) et Deligne & Bimbot (1995)); il s'agit, dans le premier cas, de déterminer quelle segmentation aboutit à la meilleure compression possible d'un corpus, en terme de sa longueur *et* de celle du lexique correspondant à la segmentation retenue (la *description* des données), sous un schéma de codage quasi optimal du genre décrit par Huffman (1952)[2]; et dans le second, de sélectionner la segmentation telle qu'elle maximise la probabilité de la génération du corpus étant donné une distribution de probabilités sur les unités du lexique.

Par contraste, la segmentation par seuillage n'implique pas la représentation explicite d'un lexique: «[algorithms belonging to this class] isolate words within a sentence only as a side-effect of correctly identifying two adjacent boundaries» (Brent 1999). En toute généralité, elle consiste en une succession de décisions ponctuelles portant sur chaque frontière potentielle dans un corpus. Soit $\sigma := s_1^n := s_1...s_n$ une séquence de n symboles pris dans l'*alphabet* $A := \{a_1,...,a_m\}$ dont les m éléments représentent typiquement les phonèmes d'une langue. Le principe est de déterminer, pour chacune des n-1 paires de symboles

1. On trouvera un résumé des principes de la théorie de la détection du signal dans Bavaud (1998), pp.142-7.

2. Ce paradigme, formalisé en premier lieu par Rissanen (1978), est connu sous le nom anglais de *minimum description* (ou *representation*) *length*.

(bigrammes) observées, s'il convient d'y insérer une frontière morphologique; la décision est basée sur l'observation d'une variable dont les valeurs élevées sont interprétées comme des signaux (de transition d'une unité morphologique vers une autre) émergeant de fluctuations non significatives.

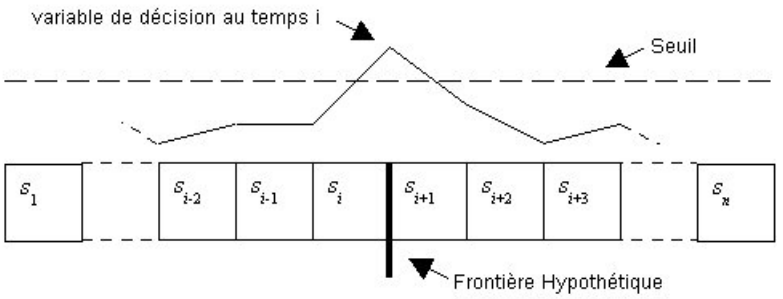

FIG. 2.1 – *Représentation schématique de la segmentation par seuillage (sur ce schéma, le seuil est fixé pour l'ensemble de la séquence)*

Formellement, il s'agit de tester pour tout i tel que $1 \leq i \leq n-1$:
$H_0(\sigma,i)$: «Le i-ème symbole de σ n'est pas suivi d'une frontière de morphème»;
$H_1(\sigma,i)$: «$H_0(\sigma,i)$ est fausse»

On rejette $H_0(\sigma,i)$ et donc on insère une frontière à la suite de s_i ssi

$$d(\sigma,i) \geq T(\sigma,i) \tag{2.1}$$

où $d(\sigma,i)$ dénote la valeur d'une *variable de décision* (ou *critère*) générique au temps i et $T(\sigma,i)$ le *seuil* arbitraire au-delà duquel la valeur de $d(\sigma,i)$ au temps i est jugée significativement élevée. Sous cette forme, le modèle est loin d'être entièrement déterminé; il ne l'est qu'après spécification de $d(\sigma,i)$ et $T(\sigma,i)$.

Dans la suite de l'exposé, cette formulation générique constituera un support unifié pour la discussion de plusieurs variables de décision envisageables. Pour en donner dès l'abord une idée plus précise, nous exposons rapidement la méthode distributionnaliste du *nombre de successeurs*, initialement décrite dans Harris (1955a), qui constitue de toute évidence la première instance attestée dans la littérature d'une procédure de segmentation par seuillage.

Dans la préface de l'édition Phoenix de «Structural Linguistics», Harris résume la version la plus simple de sa méthode dans les termes suivants:

> [Reference should be made to] a procedure for locating morpheme and word boundaries among the successive phonemes of a sentence. Given a sentence m phonemes long, for $1 \leq n \leq m$ we count after the first n phonemes of the sentence how many different $n + 1$th phonemes («successors») there are in the various sentences which begin with the same first n phonemes. If the successor count after the first n phonemes is greater both than that after the first n - 1 phonemes and than that after the first $n + 1$ phonemes of the sentence, we place a tentative morphological boundary after the nth phoneme of the given sentence. (Harris 1955b)

Le critère adopté dans cette approche est le *nombre de successeurs* (*successor count*), défini comme le nombre de symboles $a \in A$ différents pouvant apparaître à la suite du début d'énoncé s_1^i dans d'autres énoncés de la langue considérée. Harris a initialement proposé d'évaluer le nombre de successeurs en questionnant un informateur (Harris 1955a), mais il rapporte ultérieurement les résultats d'une simulation informatique où le nombre de successeurs est estimé à partir d'un corpus écrit [3] (Harris 1967). Dans ce cas, si l'on note $w \in A^i$ le i-gramme observé en s_1^i , on peut définir le nombre de successeurs comme:

$$\overrightarrow{c}(w) :=\mid \{a \in A : n(wa) > 0\} \mid \tag{2.2}$$

où $n(wa)$ dénote le nombre d'occurrences du i+1-gramme wa dans le corpus. La forme complète du test correspondant à (2.1) est donc:

$$d(\sigma,i) \geq T(\sigma,i) \begin{cases} d(\sigma,i) := \overrightarrow{c}(w) \\ T(\sigma,i) := max(d(\sigma,i-1),d(\sigma,i+1)) \end{cases} \tag{2.3}$$

où w dénote le i-gramme constitué par les i premiers symboles de l'énoncé.

Harris indique que sa méthode détecte efficacement la quasi-totalité des frontières de mots, et une part considérable des frontières de morphes. Il répartit les exceptions en deux catégories: parfois, la procédure échoue à détecter des frontières existantes (*undercuts*) parce que la distribution des successeurs en un point est *restreinte*; des exemples communs de ce genre d'erreur sont donnés par l'accord grammatical (qui restreint *à distance* le nombre de successeurs) et les alternances morphologiques conditionnées par le contexte [4]. L'autre type

3. Voir avant-propos.

4. Harris cite l'exemple anglais du signifiant *dramat-*, toujours suivi par la séquence *-ic* (Harris 1955a, p.42).

d'erreur se produit lorsque le modèle prédit à tort la présence d'une frontière (*overcuts*); ce problème est toujours imputable à un phénomène d'homonymie *partielle*:

> «The [...] most important situation is that of sectional homonyms [...], where the first part of a morpheme is identical with some whole morpheme» (Harris 1955a, p. 42).

2.2 Expression markovienne du modèle

Dans ce paragraphe, nous reformulons enterme de chaîne de Markov d'ordre k l'algorithme générique décrit au paragraphe précédent. Nous commençons par poser les hypothèses de *stationnarité* et de *conditionnement limité* - évoquant au passage la question de la forme du seuil. Puis nous dérivons successivement la *variété* et l'*entropie conditionnelle* avant d'introduire la notion de *précédence* et discuter le mode de son intégration dans le modèle[5].

2.2.1 Conditionnement de la variable de décision et forme du seuil

Comme on l'a déjà mentionné précédemment, la version du test initialement proposée par Harris fait intervenir explicitement la notion d'*énoncé*. Elle diffère donc de la formulation générique faite plus haut, où le corpus est représenté comme une séquence unique *homogène* - en ce sens qu'il n'est *pas* explicitement conçu comme une collection de sous-séquences correspondant à des énoncés indépendants[6]. Une conséquence pratique importante de cette différence est que l'estimation du nombre de successeurs (ou d'un autre indicateur parmi ceux que nous introduirons dans la suite) étant donné les $i$premiers symboles du corpus entier devient rapidement irréalisable[7] - à tout le moins considérablement plus gourmande en données d'apprentissage que dans le cas où le décompte est réinitialisé à chaque frontière d'énoncé, c'est-à-dire où i varie dans les limites du nombre de symboles effectivement observés dans les énoncés d'un corpus.

5. La plus ancienne mention faite aux hypothèses markoviennes et à l'entropie conditionnelle pour la segmentation semble être celle de Gammon (1969).

6. Dans les contextes descriptifs extrêmes que nous avons évoqués au paragraphe 1.1 (données archéologiques, cryptogrammes), l'information des frontières de phrases peut fort bien être inaccessible, ce qui justifie cette forme d'ascèse (même sans tenir compte de notre volonté explicite de promouvoir les approches non supervisées en matière de traitement automatique des langues naturelles).

7. On trouvera dans l'annexe A.3 une discussion sur la représentativité d'un modèle en fonction du corpus ayant servi à l'estimation de ses paramètres.

Pour contrer cet effet, on est amené à imposer deux contraintes sur le conditionnement de la variable de décision; ces restrictions, qui définissent le cadre théorique de la modélisation *markovienne* des séquences de symboles[8], sont connues sous le nom d'hypothèses de *conditionnement limité* et de *stationnarité*[9]. La première revient à considérer que la valeur de la variable de décision au temps i ne dépend pas *plus* de la totalité du passé connu que des k derniers symboles:

$$d(s_1^i) \approx d(s_{i-(k-1)}^i) \tag{2.4}$$

La seconde exprime l'idée que la valeur de la variable de décision ne dépend pas explicitement du temps i:

$$(\forall i) d(s_{1+i}^{k+1}) \approx d(s_1^k) \tag{2.5}$$

En conjuguant les deux hypothèses (dites *markoviennes*), on aboutit à une quantité de la forme $d(w)$, soit uniquement conditionnée par la séquence de k symboles (ou k-gramme) $w \in A^k$ se terminant à la position considérée, et entièrement définie par m^k paramètres (rappelons que m désigne la taille de l'alphabet A et A^k l'ensemble des combinaisons de k symboles pris dans A).

Notons que Harris argumente dès le premier article en défaveur des simplifications en question[10]. Toutefois, comme nous le verrons dans la suite, il est possible récupérer une part considérable des perturbations induites par l'observation de séquences courtes en complexifiant le modèle par l'introduction des extensions qui font l'objet des alinéas 2.2.3.2 et 2.2.3.3.

L'expression définitive du test (2.1), implique également de déterminer la relation spécifique liant la valeur du seuil $T(\sigma,i)$ à ses arguments σ et i. La procédure originale de Harris maximise *localement* le nombre de successeurs, c'est-à-dire qu'elle insère une frontière à chaque position s_i où la variable de décision est à la fois plus élevée qu'aux deux positions adjacentes s_{i-1} et s_{i+1}. Cette spécification de $T(\sigma,i)$ traduit le présupposé que la probabilité d'insérer une frontière en un point donné de la séquence *dépend* du comportement du critère immédiatement avant et après la position examinée. Ce raisonnement est intuitivement justifiable, mais on peut aussi considérer que, puisque le critère $d(\sigma,i)$ est toujours[11] conditionné par un k-gramme $w \in A^k$ et prend donc ses

8. Voir annexe A.1

9. Voir par exemple Manning & Schütze (1999, p.318); on trouvera une autre formulation des deux hypothèses dans l'annexe A.1, expressions (A.1) et (A.2).

10. Voir la citation faite dans la note 6, p. xvi.

11. du moins sous les hypothèses (2.4) et (2.5)

valeurs dans un intervalle fixé pour l'ensemble de la séquence [12], le seuil au-delà duquel le signal est tenu pour significativement élevé devrait être constant et conditionné uniquement sur $\sigma : T(\sigma,i) \approx T(\sigma)$; le problème est alors de trouver une *bonne* valeur de seuil, telle qu'elle conduit à la segmentation souhaitée dans la majeure partie des cas considérés.

Il serait erroné de croire que la conception de Harris exclut ce type de décision arbitraire portant sur une valeur absolue; en effet, le test (2.1) ne spécifie pas la décision à prendre lorsque deux ou plusieurs symboles consécutifs obtiennent une valeur constante; c'est ce qui amène Harris à noter que:

> [...] segmental morphemes which consist of one phoneme are not easily separated out, since their boundary may be overshadowed by the neighboring boundary. In any case, a plateau of two high numbers [...] indicates two segmentations, even though there are not two separate peaks. (Harris 1955a, p.67)

Si l'on omet cette correction, le test est tout bonnement incapable de détecter correctement les frontières de signifiants constitués par un symbole unique. Même ainsi, le traitement de ce problème n'est pas complètement déterminé, puisqu'il implique encore de fixer le niveau au-delà duquel des valeurs successives égales (formant *plateau*) sont considérées comme *élevées*. En outre, Harris ne précise pas les raisons qui l'amènent à inférer la présence de frontières aux points où le critère est localement *maximal* - plutôt que *plus élevé* qu'aux positions adjacentes *d'au moins* x, ou encore localement maximal *et supérieur à* x [13].

Dans l'état actuel de notre réflexion sur la question, nous n'avons pas les moyens d'argumenter plus avant pour l'une ou l'autre approche. Ce sont en définitive des critères de simplicité algorithmique qui nous conduisent à adopter dans ce travail la conception d'un seuil constant d'un bout à l'autre de la séquence. C'est sans doute l'un des aspects du modèle à propos duquel le plus d'encre doit encore couler, ainsi que nous le répéterons dans la partie 4.

2.2.2 Critères

Dans cet alinéa, nous exposerons successivement les indicateurs susceptibles de jouer le rôle de variable de décision dans la formulation générique (2.1). Cette présentation sera divisée en trois parties. Nous dériverons tout d'abord

12. par exemple, dans le cas du nombre de successeurs, il est évident qu'il ne sera jamais supérieur à la taille m de l'alphabet (ni inférieur à zéro)

13. x dénote un second paramètre dont l'unité dépend du critère sélectionné

la *variété conditionnelle* [14], c'est-à-dire l'homologue, sous les hypothèses de stationnarité et de conditionnement limité [15], du *nombre de successeurs* défini par Harris (1955a). Nous tenterons ensuite de justifier deux extensions qui améliorent les performances du système sous les hypothèses en question: le passage de la variété à l'*entropie conditionnelle*, et l'intégration de la notion que nous proposons d'appeler *précédence*.

2.2.2.1 Variété Conditionnelle

La *variété* conditionnelle étant donné un k-gramme $w \in A^k$ est définie de la même façon que le nombre de successeurs (2.2), à ceci près que la taille du contexte w considéré est constante est égale à k plutôt que croissante et égale à i:

$$\overrightarrow{v}(w) :=| \{a \in A : n(wa) > 0\} | \, , w \in A^k \tag{2.6}$$

Il en résulte une diminution importante du nombre de paramètres du modèle - qui correspond à la quantité d'information *syntaxique* [16] à laquelle on renonce de cette façon. Cette quantité est d'autant moindre que k est élevé, mais elle n'est jamais négligeable, comme le confirment les résultats présentés dans la partie 4. Dans les conditions où la méthode de Harris donnerait une segmentation très *exacte* [17], les hypothèses markoviennes induisent inévitablement un *bruit* considérable. C'est dans ce cadre *approximatif* que devront être appréciées les performances obtenues, et c'est l'une de nos principales préoccupations dans ce travail que de questionner certains aspects d'une méthodologie de description linguistique «en eaux troubles». A ce titre, les extensions discutées dans les alinéas suivants peuvent être conçues, métaphoriquement, comme des paramètres dont le réglage vise à améliorer la discrimination entre un *signal* morphologique affaibli [18] et des *perturbations* relevant d'autres niveaux de la structure de la langue [19].

14. à distinguer, rappelons-le, de la *variété* au sens de Harris (voir note 9, p. xviii)

15. Voir alinéa précédent.

16. au sens propre de «caractérisant les relations entre signes morphologiques au niveau de l'énoncé»

17. encore qu'elle ne soit pas conçue comme un substitut de l'*analyse* morphologique proprement dite: «morphological procedures which describe the structure of the utterances as a distribution of these segments (and in doing so correct the segmentation to obtain better structural elements)» (Harris 1955a, p.33)

18. par l'usage des hypothèses de stationnarité et de conditionnement limité

19. Notons que le caractère *perturbateur* des contraintes d'ordre syllabique, syntaxique ou sémantique qui influencent la succession des symboles n'existe que dans la mesure où elles

2.2.2.2 Standardisation: entropie conditionnelle

Nous avons vu dans l'alinéa précédent que la variété conditionnelle est déterminée par la *présence* ou l'*absence* des k+1-grammes wa ($a \in A$,$w \in A^k$ fixé) dans le corpus examiné. Dans cette perspective, l'information de la *fréquence* de wa n'est pas prise en compte; comme l'indique Harris:

> In general, frequency of occurrence correlates with what may be considered language USE (or communication) as against language STRUCTURE; beyond this point the investigations will ask whether a sound [...] EVER occurs in a given environment [...] rather than how frequently it occurs there. (Harris 1955, p.59, souligné par l'auteur)

Le présupposé qui sous-tend cette position est que l'*usage* n'est pas un bon indicateur de la *structure*, ou du moins qu'il n'est pas utile à son inférence. De fait, des contraintes de cohérence thématique et de cohésion stylistique peuvent altérer profondément les propriétés statistiques des séquences de symboles. Par exemple, dans le présent document, la forme *avons* est beaucoup plus fréquente que *avoir*. Du point de vue structurel, pourtant, la frontière qui sépare la racine *av-* du suffixe *-ons* ne diffère en rien de celle qui la sépare du suffixe *-oir*. La différence perçue au niveau de l'usage ne correspond pas à une distinction pertinente au niveau de la structure.

Nous pensons toutefois que le *bruit* induit par les hypothèses markoviennes justifie qu'on tienne compte des fréquences; il faut en effet considérer que, dans notre perspective, *av* peut être suivi par *ant*, *aud*, *erser*, etc. (*avant*, *Bavaud*, *traverser*), et il est clair qu'on ne souhaite pas que ces successeurs accroissent la variété conditionnelle - ou, plus généralement, la conviction du système de la présence d'une frontière - de la même façon que les précédents. Comme notre modèle spécifie une valeur unique pour un k-gramme donné, nous savons qu'en pareil cas, seule une part des occurrences sera traitée correctement, c'est-à-dire qu'il détectera une frontière de morphe après *av* dans *av-ons* et *av-oir ou* n'en détectera pas dans *avant*, *Bavaud*, *travers*(*-er*). En dernière analyse, le critère de sélection est la fréquence des unités: si *avons*, *avoir* et les formes comparables sont plus fréquentes qu'*avant*, *Bavaud*, *traverser* et les formes comparables, l'option engendrant le moins d'erreur est d'insérer systématiquement une frontière

---

peuvent être de *mauvaises explications* du niveau de la structure linguistique que l'on cherche à mettre en évidence, ce qui n'est probablement pas le cas général dans les langues naturelles; l'*accord* est un exemple fréquent de trait syntaxique *perturbant* à distance le nombre de successeurs (voir Harris 1955a, p.42 et pp.61-62), et dont l'influence est paradoxalement corrigée par la variété conditionnelle, qui rompt les dépendances entre des éléments séparés par plus de k symboles.

après *av*.

Si l'on admet cet argument, la question qui se pose est de savoir *de quelle façon* l'information fréquentielle peut être intégrée au modèle. Une première réponse s'appuie sur la notion de *probabilité de transition*. Rappelons qu'une chaîne de Markov est définie par la donnée de l'ordre k et de m^{k+1} probabilités de transitions $p(w \rightarrow a), w \in A^k, a \in A$ (voir annexe A.1). Pour un contexte $w \in A^k$ donné, la variété conditionnelle est liée aux probabilités de transition par:

$$\overrightarrow{v}(w) :=\mid \{a \in A | p(w \rightarrow a) > 0\} \mid \tag{2.7}$$

L'information conditionnelle $i(w \rightarrow a) := -\log p(w \rightarrow a)$ (A.8) peut être interprétée comme une mesure de la surprise associée à la transition $w \rightarrow a$. Dans une perspective morphologique, on peut s'attendre à ce que les transitions de symboles se produisant à l'intérieur d'unités morphologiques soient généralement moins surprenantes que les transitions entre unités [20]. A ce titre, $i(w \rightarrow a)$ constitue un candidat intéressant pour jouer le rôle de variable de décision dans (2.1), qui se réécrit alors:

$$i(w \rightarrow a) > T(\sigma) \qquad w = s_{i-(k-1)}^{i}, a = s_{i+1} \tag{2.8}$$

Une différence fondamentale entre variété et information conditionnelle réside dans le conditionnement de la seconde sur w *et* a. A l'inverse, $\overrightarrow{v}(w)$ est une mesure globale sur *tous* les successeurs de w. Pour un ordre k donné, le nombre de paramètres du modèle basé sur $P(w \rightarrow a)$ est m fois plus élevé que celui du modèle basé sur $\overrightarrow{v}(w)$. L'interrogation suivante paraît légitime à ce point de l'exposé: pour un contexte $w \in A^k$ donné, est-il possible de combiner la surprise associée à chaque transition $w \rightarrow a, a \in A$ dans un indicateur ne dépendant plus de a, donc conditionné uniquement sur w? En d'autres termes, peut-on trouver un compromis entre la concision de la variété et la précision de l'information?

Il est bien au-delà des ambitions de ce travail que de démontrer le résultat suivant, dont on trouvera la preuve la plus élégante dans Shannon (1948): le seul indicateur présentant toutes les qualités souhaitables [21] pour une mesure de l'incertitude associée aux successeurs possibles d'un contexte $w \in A^k$ est l'*entropie*

20. C'est notamment la perspective adoptée par Saffran, Newport & Aslin (1996); leur recherche porte sur les transitions de syllabes, mais ainsi que les auteurs le soulignent: «Whether statistics are computed across phonemes, syllables, or other subword units [...], the same types of learning mechanisms should suffice in principle for the induction of word boundaries» (Saffran, Newport & Aslin 1996, p.611).

21. On trouvera dans l'annexe A.2 une discussion plus approfondie des propriétés de l'entropie.

conditionnelle étant donné w (A.15) définie comme l'espérance mathématique de l'information associée à chaque transition partant de w: $\overrightarrow{h}(w) := E(i(w \rightarrow a))$ w fixé. Typiquement, on s'attend à ce que l'entropie conditionnelle, qui varie entre 0 (déterminisme) et log m (équiprobabilité), soit généralement plus élevée pour les k-grammes précédant une frontière morphologique (Hutchens & Alder 1998).

Nous verrons dans la partie 4 que, sur des corpus corrompus par la suppression des frontières d'énoncés (et pour une tâche de segmentation en mots), l'entropie conditionnelle donne systématiquement de meilleurs résultats pour la segmentation en *mots* que la variété correspondante. Il semble ainsi qu'en tenant compte des fréquences, on puisse souvent obtenir une meilleure discrimination entre les régularités pertinentes observées au niveau de l'énoncé et celles - *accidentelles*, qui résultent de la concaténation des énoncés.

2.2.2.3 Transition et précédence

Les quantités discutées précédemment ont une faiblesse en commun: elles ne tiennent compte que des dépendances séquentielles *de gauche à droite*. Il en résulte un fait que nous avons pris la liberté d'occulter jusqu'ici: en basant la segmentation sur la variété ou l'entropie conditionnelle telles que nous les avons définies plus haut, nous n'identifions pas exactement les *frontières* d'unités morphologiques mais plutôt les *fins* d'unités.

Ce constat fait l'objet d'une discussion détaillée de la part de Harris, qui envisage le *nombre de prédécesseurs* comme une correction à appliquer aux résultats du nombre de successeurs:

> The backward operation is then no closer an approximation to morpheme boundaries than is the forward; but it is a check on the forward operation (Harris 1955a, p.43).

Il ne donne pas d'indication plus précise sur les modalités de la combinaison des deux tests, mais signale qu'elle permet de régler la plupart des problèmes évoqués au paragraphe 2.1, notamment tous les cas où l'erreur n'est induite que par l'une des deux lectures; dans ces circonstances, un pic erroné (ou une absence de pic erronée) sur l'une des courbes est toujours contrebalancé(e) par une absence de pic correcte (ou un pic correct) sur l'autre[22].

22. Le problème est qu'on ne dispose pas, à ce niveau, d'un moyen de faire le départ entre les deux situations: «'The decison among these possibilities can be made only by morphological tests [...]» (Harris 1955a, p.42)

Revenant à notre formalisme, il semble donc justifié de reprendre le raisonnement au début en définissant la *probabilité de précédence*[23] $p(a \leftarrow w)$ qu'un contexte w soit *précédé* d'un symbole a. Nous parlerons désormais de chaîne de Markov *inverse* (d'ordre k), définie par une *matrice de précédence* $\overrightarrow{P}$ de composantes $\overrightarrow{P}_{wa} := p(a \leftarrow w) = \frac{p(aw)}{p(w)}$. La variété conditionnelle *inverse* peut alors être définie comme:

$$\overleftarrow{v}(w) :=\mid \{a \in A | p(w \leftarrow a) > 0\} \mid \tag{2.9}$$

L'entropie conditionnelle *inverse*[24] est définie à partir de l'information conditionnelle *inverse*:

$$\overleftarrow{h}(w) := E_a(i(w \leftarrow a)) = -\sum_{a \in A} p(w \leftarrow a) \log p(w \leftarrow a) \tag{2.10}$$

et dénote l'incertitude sur la distribution des symboles pouvant *précéder* w.

Si l'on utilise l'un ou l'autre de ces indicateurs pour tester la présence d'une frontière entre les i-ème et i+1-ème symboles d'une séquence, il convient de définir leurs arguments de façon telle que $a = s_i$ et $w = s_{i+1}^{i+k}$. En général, s'attend à ce qu'un système basé sur une quantité inverse (plutôt que son homologue *standard*) soit un meilleur outil pour la détection des *débuts* d'unités morphologiques.

Du point de vue statistique, l'une des qualités intéressantes des modèles basés sur les probabilités de précédence (et les quantités dérivées) est que leurs paramètres peuvent être estimés à partir du même décompte de k+1-grammes que les modèles transitionnels correspondants; en matière de représentativité, bien sûr, des contraintes identiques s'appliquent dans les deux cas (voir annexe A.3).

Il est possible d'évaluer simultanément la probabilité qu'un point de la séquence marque la fin d'une unité *et* le début d'une autre en additionnant entropie conditionnelle standard et inverse. Plus rigoureusement, si l'on définit $w = s_{i-k+1}^{i}$, $w' = s_{i+1}^{i+k}$, $a = s_{i+1}$ et $a' = s_i$ la présence d'une frontière après s^i

23. Nous ne connaissons pas de terme conventionnel pour désigner cette notion en français; si elle n'a pas encore été baptisée, ce néologisme en vaut bien un autre.

24. Notons que la convention retenue dans ce travail d'apposer le qualificatif *inverse* au nom de l'indicateur pour marquer l'opposition entre quantités transitionnelles et précédentielles induit une ambiguïté fâcheuse: par exemple, nous parlerons d'entropie conditionnelle pour désigner à la fois le type d'indicateur (par opposition à la variété conditionnelle) et la direction (par opposition à l'entropie conditionnelle *inverse*); c'est ce qui nous conduira quelquefois à utiliser le terme *standard* comme antonyme explicite d'*inverse*.

est d'autant plus vraisemblable que $\overrightarrow{h}(w)$ et $\overleftarrow{h}(w')$ sont élevées; pour exprimer la combinaison dans la même unité que ses parties, on peut encore diviser la somme par deux, obtenant ainsi la moyenne arithmétique, que nous convenons d'indiquer par une barre horizontale:

$$\overline{h} = \frac{\overrightarrow{h}(w) + \overleftarrow{h}(w')}{2} \tag{2.11}$$

Notons que rien n'indique *a priori* que les composantes transitionnelle et précédentielle doivent peser le même poids sur le résultat. On peut tester d'autres pondérations en introduisant un *coefficient de transition* $0 \leq \overrightarrow{c} \leq 1$ mesurant l'importance accordée à la mesure transitionnelle; on dira alors que la présence d'une frontière est d'autant plus vraisemblable que $\overline{h}(w,w',\overrightarrow{c}) := \overrightarrow{c}\,\overrightarrow{h}(w) + (1-\overrightarrow{c})\,\overleftarrow{h}(w')$ est élevée pour généraliser (2.11). Dans cette formulation, $\overrightarrow{h}(w)$, $\overleftarrow{h}(w)$ et $\overline{h}(w,w')$ correspondent respectivement aux cas $\overrightarrow{c} = 1, \overrightarrow{c} = 0$ et $\overrightarrow{c} = 0.5$.

Ainsi qu'on le verra dans la partie 4, les résultats de notre expérimentation indiquent que l'utilisation de la combinaison (2.11) favorise la segmentation en *mots* au détriment du niveau *morphique*. A première vue, cela semble contredire la conception de Harris, pour qui le nombre de prédécesseurs permet justement d'améliorer la segmentation en morphes à l'intérieur des mots en révélant des pics de diversité absents de la courbe des successeurs (et vice versa). Mais il importe de considérer que sa méthode diffère de celle exposée ici notamment par la forme du seuil[25], qui porte dans notre cas sur les *valeurs élevées* et non les *pics*. En combinant les deux indicateurs sous forme de moyenne, on peut s'attendre à ce que les frontières morphologiques apparaissant à l'intérieur de mots obtiennent généralement un score plus faible que les frontières entre mots[26], d'où une discrimination plus marquée à ce niveau.

25. Voir alinéa 2.2.1

26. quoique plus élevé que les transitions qui ne sont pas des frontières du tout

Chapitre 3

De l'entropie conditionnelle à l'information mutuelle

Dans la partie précédente, nous avons cherché à traduire en termes markoviens la méthode du nombre de successeurs de Harris. Nous avons ainsi donné une formulation explicite du test générique qui sous-tend la procédure, avant de d'exposer deux variables de décision plausibles dans ce contexte, la variété et l'entropie conditionnelle. Enfin, nous avons discuté l'intérêt des quantités précédentielles pour la segmentation, et les modalités formelles de leur intégration dans le modèle.

Nous considérons maintenant la généralisation du formalisme dans le sens dont nous avons cherché à donner l'intuition dans l'avant-propos et le paragraphe 1.3, c'est-à-dire qu'on s'efforcera essentiellement à justifier:

a. la substitution d'un critère de diversité (sur la distribution conditionnelle du contexte examiné) par un critère de (*dis-*)*similarité* (entre cette distribution et celle, dite *initiale*, attendue en début d'unité[1]), en particulier la *divergence de Kullback-Leibler* ou entropie *relative*;

b. l'approximation de la distribution initiale, généralement inconnue, par la distribution, dite *inconditionnelle*, des symboles *hors contexte*, ce qui revient à utiliser comme variable de décision l'*information mutuelle moyenne* entre le contexte et ses successeurs, qui est une mesure de la contrainte qu'il impose à leur distribution.

1. ou en fin d'unité, si l'on utilise un indicateur *inverse* (au sens défini dans l'alinéa précédent), auquel cas on parlera de distribution *finale*; notons qu'afin d'éviter d'accroître encore la redondance de ce texte, nous baserons notre discussion uniquement sur le cas transitionnel.

Ces réflexions achèveront de poser le cadre théorique de ce travail, dont la partie suivante sera consacrée à la présentation des résultats expérimentaux obtenus sur deux corpus orthographiques.

3.1 Incertitude et divergence

Jusqu'ici, nous avons modélisé la segmentation en terme de seuillage d'une mesure d'incertitude ou, plus généralement, de diversité sur une distribution conditionnelle. Nous n'avons pas fait plus, à ce point, que de décliner la méthode du nombre de successeurs sur un mode markovien. C'est un mouvement théorique plus fondamental que nous décrivons ici, puisqu'il aboutit à l'introduction d'une seconde distribution, celle des symboles *en début d'unité* ou distribution *initiale*, dont il s'agit d'évaluer combien elle diffère de la première; si cette différence est faible, il est vraisemblable que le contexte considéré soit effectivement une fin d'unité. Notons que rien en cette procédure ne devrait surprendre un lecteur de Harris, qui, qualifiant la méthode du nombre de successeurs, indique clairement que:

> This is a special case, though the most common one. More generally: we segment the utterance at those points where the number and variety of successors [...] is similar to that at utterance end. (Harris 1955a, p.65)

Au premier abord, cette approche semble n'avoir que peu de traits communs avec celle que nous avons discutée précédemment. Pourtant, leurs expressions formelles sont étonnamment proches, et leur comparaison permet de comprendre que la procédure basée sur la diversité est bel et bien un cas particulier de celle basée sur la (dis-)similarité, où l'on postule implicitement que la distribution initiale est uniforme.

Rappelons que le test de la présence d'une frontière sur la base de l'entropie conditionnelle est formulé comme:

$$\overrightarrow{h}(w) > T(\sigma) \qquad w := s_{i-(k-1)}^{i} \tag{3.1}$$

et que $\overrightarrow{h}(w) := E_a(i(w \rightarrow a)) = -\sum_a p(w \leftarrow a) \log p(w \leftarrow a)$, où la sommation est effectuée sur tous les symboles $a \in A$ de probabilité non-nulle. Soit $\widetilde{p} := \widetilde{p}(a_1), \ldots, \widetilde{p}(a_m)$ la distribution *initiale* des symboles et $p^w := p(w \rightarrow a_1), \ldots, p(w \rightarrow a_m)$ la distribution conditionnelle étant donné $w \in A^k$, qui toutes deux portent sur l'alphabet A. La mesure de dissimilarité la plus utilisée dans la cadre de la Théorie de l'Information est la *divergence* (*de Kullback-Leibler*) ou entropie *relative* (A.22) entre deux distributions sur le même alphabet, définie dans notre cas comme:

$$K(p^w \parallel \widetilde{p}) := \sum_a p(w \to a) \log \frac{p(w \to a)}{\widetilde{p}(a)} \tag{3.2}$$

La divergence possède en particulier la propriété de s'annuler ssi la probabilité de chaque symbole est la même dans les deux distributions, ce qui nous permet de la substituer à l'entropie conditionnelle dans le test (3.1) - en inversant l'inégalité, puisqu'il s'agit de *minimiser* la dissimilarité. En outre, les deux hypothèses (rendues nécessaires par le caractère logarithmique de l'expression) $\log \frac{0}{\widetilde{p}(a)} = 0$ et $\log \frac{p(w \to a)}{0} = \infty$

ont des implications moins fondamentales mais tout de même appréciables: la première indique que la divergence ne prend en compte que les symboles observés dans le contexte considéré (et non les symboles autorisés en position initiale mais non observés dans le contexte), ce qui révèle la *tolérance* de la mesure [2]; à l'inverse, la seconde exige que la divergence soit infinie lorsque l'un au moins des symboles interdits à l'initiale est attesté dans la position examinée - ce qui est naturel puisque la distribution initiale est supposée *vraie*.

Comme on le voit, dans un formalisme markovien, le passage de la variété à la dissimilarité est simple et direct; concrètement, il se traduit par le rapport des probabilités de transition aux probabilités initiales, et ne nécessite donc que de connaître en plus les m paramètres de la seconde distribution. Nous verrons dans la partie suivante que cette petite addition améliore considérablement les performances de la méthode, parfois autant que la multiplication qui résulterait du passage à un ordre supérieur. En revanche, son défaut incurable est de nous faire sortir d'un cadre strictement *non-supervisé*; si nous connaissions la distribution initiale, il semble que la raison d'être de toute notre construction serait remise en cause, et nous avons suffisamment répété que ce n'était pas le cas.

Dès lors, le vrai problème est d'acquérir les paramètres de cette distribution à partir d'un corpus où les traces formelles qui nous permettraient de le faire ne figurent pas. Si une première segmentation du corpus est disponible [3], on peut faire l'hypothèse que les séquences de symboles caractérisant les fins de segments devraient avoir la même fonction de *rupture* lorsqu'elles se produisent à l'intérieur d'un segment - ce qui correspond à une forme d'apprentissage par *généralisation* [4]. Mais si, comme on l'admet dans ce mémoire, la séquence ne contient aucune forme de séparateur *a priori*, on est contraint recourir à d'autres hypothèses. C'est ce qu'on a fait implicitement dans la partie précédente en

2. en ce sens qu'elle tolère la possible imperfection de la distribution conditionnelle
3. comme les frontières d'énoncés dans le cas de Harris (1955)
4. Voir note 6, p. xvi

utilisant l'entropie conditionnelle; dans ce cas, l'hypothèse est que la distribution initiale est uniforme, d'où:

$$
\begin{aligned}
&K(p^w \parallel \widetilde{p}) := \sum_a p(w \to a) \log \frac{p(w \to a)}{1/m} \\
&= \sum_a p(w \to a) \log(m\, p(w \to a)) \\
&= \sum_a p(w \to a) \log m + \sum_a p(w \to a) \log p(w \to a) \\
&= \log m - \overrightarrow{h}(w)
\end{aligned}
\tag{3.3}
$$

Comme le terme $\log\, m$ est constant, on peut le négliger dans le cadre du seuillage; il subsiste $-\overrightarrow{h}(w)$, dont le test (3.1) revient à maximiser l'inverse.

Mais cette simplification n'est pas toujours praticable; les résultats discutés dans la partie 4 indiquent notamment qu'elle est moins efficace dans le cas du latin que de l'anglais, ce qui signifie que les *bons* débuts et les *bonnes* fins d'unités dans la première langue sont moins régulièrement caractérisées par une entropie conditionnelle élevée que dans la seconde. Quelle meilleure approximation de la distribution initiale pourrait-on suggérer dans ce cas?

3.2 Divergence et information mutuelle

Pour aborder la dernière ligne droite dette présentation théorique, nous proposons de formuler la question différemment: pourquoi postuler que la distribution initiale $\widetilde{p}$ est uniforme alors que la distribution (dite *inconditionnelle*) des symboles observés dans le corpus p[1] ne l'est généralement pas? Si l'on a saisi le sens de la nuance, ce qui suit ne fera que confirmer l'intuition.

En calcul élémentaire des probabilités, deux événements A et B sont dits *indépendants* ssi $p(A \cap B) = p(A)p(B)$. Dans le formalisme adopté ici, on dira plutôt que le contexte w et son successeur a sont indépendants ssi $p(wa) = p(w)p(a)$; dans ce cas (et seulement dans ce cas), on a[5]:

$$p(w \to a) = \frac{p(wa)}{p(w)} = \frac{p(w)p(a)}{p(w)} = p(a) \tag{3.4}$$

5. Voir par exemple Bavaud (1998), p.19.

On parle de dépendance *positive* ssi $p(w \rightarrow a) > p(a)$, c'est-à-dire si le symbole a est plus probable quand il est précédé du contexte w, et de dépendance *négative* dans le cas contraire $p(w \rightarrow a) < p(a)$. Dans une perspective morphologique, on peut s'attendre à ce que w et a soient dans un rapport de dépendance positive si la transition $w \rightarrow a$ se produit à l'intérieur d'une unité, et dans un rapport d'indépendance (ou de dépendance négative) sinon. En d'autres termes, les points de segmentation les plus probables ne sont pas simplement ceux où l'on est surpris du déroulement de la séquence[6], mais ceux où l'on est surpris alors même que le symbole apparu est fréquent. On peut donc réécrire (3.1) comme:

$$\frac{p(w \rightarrow a)}{p(a)} \leq T(\sigma) \qquad w = s_{i-(k-1)}^{i}, a = s_{i+1} \tag{3.5}$$

Comme le logarithme est une fonction strictement croissante, on obtient le même résultat en utilisant *l'information mutuelle ponctuelle*[7] associée à une transition $w \rightarrow a$ (A.9), définie par $I(w \rightarrow a) := -\log \frac{(w \rightarrow a)}{p(a)}$. Cette mesure est positive ssi a est plus probable après w qu'il ne l'est dans l'absolu, nulle ssi la probabilité de a n'est pas influencée par la présence de w et négative ssi a est moins probable après w.

Ainsi qu'on l'a fait pour l'information conditionnelle, il semble raisonnable de s'interroger sur la possibilité de déterminer, pour un contexte w donné, l'information mutuelle *moyenne* associée à la distribution de ses successeurs. Cette quantité existe et se révèle très similaire, dans sa définition, à l'entropie conditionnelle[8]:

$$\overrightarrow{I}(w) := E_a(I(w \rightarrow a)) = \sum_a p(w \rightarrow a) \log \frac{p(w \rightarrow a)}{p(a)} \tag{3.6}$$

Au signe près, la différence entre $\overrightarrow{h}(w)$ et $\overrightarrow{I}(w)$ tient à la présence du diviseur $p(a)$ dans la seconde. L'information mutuelle moyenne est toujours positive ou nulle (voir par exemple Cover & Thomas 1991), et s'annule ssi les successeurs de w sont *indépendants* de lui, c'est-à-dire si la distribution des symboles *étant donné* w est identique à la distribution inconditionnelle. Que le lecteur ait ou non fait de lui-même le lien avec la divergence (3.2) n'enlève rien à sa force:

6. comme c'est le cas dans l'approche basée sur l'information conditionnelle (voir alinéa 2.2.2.2)

7. Voir Brent (1999) pour une première application au problème de la segmentation; on trouvera une présentation plus théorique dans Manning & Schütze (1999), p.68.

8. Voir annexe A.2, équation (A.19).

fonder la segmentation sur l'indépendance des successeurs du contexte examiné ou sur la divergence entre distribution conditionnelle et inconditionnelle sont une seule et même pratique.

A notre connaissance, l'usage de l'information mutuelle moyenne n'est mentionné dans aucune publication portant sur la segmentation morphologique. Nous verrons pourtant dans la partie suivante que, dans les conditions de notre expérience, ses performances sont au pire un peu moins bonnes, et au mieux nettement meilleures que celles de l'entropie conditionnelle. Dans le cas du latin, elle conduit même à une meilleure segmentation que la divergence entre la distribution conditionnelle et la *vraie* distribution initiale (telle qu'estimée dans le corpus avant la suppression des frontières), en particulier aux ordres les plus bas.

Chapitre 4

Expérimentation

Dans cette partie, nous décrivons les conditions et résultats d'une tentative d'application des concepts développés dans les parties 2 et 3 à deux corpus orthographiques anglais et latin. Nous commencerons par rappeler les hypothèses que nous cherchons à mettre à l'épreuve, et proposerons une procédure d'expérimentation visant à dégager des indices de leur validité. Ensuite, nous discuterons plusieurs problèmes relatifs à la sélection des corpus et à la norme de codage observée. Puis nous introduirons brièvement les mesures qui nous permettront d'évaluer objectivement les performances du modèle [1] en fonction de variations paramétriques. Enfin, nous commenterons les résultats proprement dits en tâchant de mettre en exergue ceux qui concernent le plus directement nos hypothèses.

4.1 Hypothèses et procédure

Ainsi qu'on l'a évoqué à plusieurs reprises dans les parties précédentes, l'expression du critère de segmentation de Harris en terme d'incertitude, de divergence et d'indépendance est fondée par un ensemble d'hypothèses, dont les plus importantes sont sans doute les suivantes:

a. il devrait être possible d'améliorer passablement les résultats si l'on connaît la *vraie* distribution *initiale* et que l'on cherche à minimiser la divergence entre cette distribution et la distribution conditionnelle;

1. en l'occurrence pour la segmentation en mots

b. on devrait pouvoir remplacer avantageusement l'hypothèse que la distribution initiale est uniforme par celle qu'elle est la même que la distribution *inconditionnelle*, en particulier dans les langues comme le latin où la diversité explique moins bien la segmentation.

L'objectif principal de notre expérimentation est de mettre la validité ces hypothèses à l'épreuve des données. Dans les grandes lignes, les étapes de la démarche sont classiques: sélection des paramètres variés, sélection et codage des corpus, définition d'une mesure d'évaluation, enfin production et interprétation des résultats. La question des paramètres sera discutée dans un instant, et les autres aspects de la procédure feront l'objet des paragraphes suivants.

La plus importante variation considérée ici est celle du *critère* utilisé. Il est clair qu'on ne peut rien conclure pour nos hypothèses si l'on ne peut pas comparer les résultats obtenus sur les mêmes données par la variété et l'entropie conditionnelle, la divergence et l'information mutuelle[2]. La variation de la *langue* concerne plus spécifiquement l'hypothèse b), qui porte sur la généralité des critères considérés. Comme nous le verrons dans le paragraphe suivant, ce sont des corpus «proches» de l'anglais et du latin (écrits) que nous ferons contraster dans cette perspective. Nous chercherons également à faire varier l'*ordre* du modèle utilisé, mais de façon moins systématique, puisque aucune de nos hypothèses n'est explicitement concernée par cette dimension.

4.2 Echantillonnage et codage

Les corpus retenus pour cette expérimentation l'ont été sur la base de plusieurs critères. Le premier et plus contraignant est leur disponibilité sur le web[3] (aux côtés de nombreux autres). Les suivants sont plus pertinents pour notre propos. En particulier, on s'est efforcé de trouver deux corpus satisfaisant les conditions suivantes:

a. ils sont rédigés dans deux langues dont on peut s'attendre à ce qu'elles induisent un comportement différent de l'entropie conditionnelle;

2. Il importe de signaler que, dans la suite, nous utiliserons *systématiquement* les quantités *moyennes* (au sens défini dans l'alinéa 2.2.2.3) avec un coefficient de transition constant $\vec{c} = 0.5$, ce qui correspond à l'équation (2.11).

3. En l'occurrence, dans l'excellente collection de textes électroniques maintenue sur le serveur de l'Université de Virginie: http://etext.lib.virginia.edu/ .

b. les langues en question nous sont suffisamment connues pour que nous puissions interpréter les performances des indicateurs;

c. le codage conserve une trace des frontières visées (en l'occurrence celles des mots), afin qu'on puisse pratiquer une évaluation systématique;

d. les corpus sont assez longs pour que les estimations soient représentatives;

e. ils sont codés sur des alphabets plutôt réduits, afin de parer à une croissance trop rapide du nombre de paramètres du modèle.

Il nous a semblé que le roman «Emma» de Jane Austen (en anglais) et les «Métamorphoses» d'Ovide (en latin) correspondaient assez bien à ces exigences. En particulier, le latin est une langue où la suffixation (flexionnelle) est un phénomène banal, notamment par rapport à l'anglais, où elle s'avère plutôt rare[4]. On peut donc s'attendre à ce qu'elle pose plus de problèmes pour la segmentation en *mots*, puisqu'on sait des affixes (appelés par Harris morphèmes *liés*) que, s'ils ont souvent une distribution plus restreinte que les mots[5], ils sont très généralement caractérisés par une plus grande liberté distributionnelle que les séquences de symboles qui ne sont pas des fins ou des débuts de signifiants de morphes.

Les deux textes sont codés sur des alphabets raisonnablement petits: 26 en anglais et 23 en latin[6] après suppression de tous les signes non alphabétiques[7]. Nous avons d'abord choisi de ne conduire l'expérience que pour les ordres 2 et 3, limites fixées vers le bas par la conviction qu'un modèle d'ordre 1 est insuffisant, et vers le haut par la complexité rapidement démesurée des modèles d'ordre supérieur - en plus de la difficulté de trouver des corpus de taille suffisante à leur estimation. Comme on sait par (A.27) que les paramètres d'un modèle

4. Plus précisément, la suffixation en anglais porte sur des paradigmes très restreint (moins de cinq successeurs en moyenne, disons), ce qui a pour effet d'accroître en général la probabilité des formes non-marquées (contenant un suffixe zéro) par rapport au latin; la suffixation en anglais n'est pas *rare* mais *rarement marquée formellement.*

5. Rappelons que le mot est défini par Bloomfield (1933) comme *forme libre minimale.*

6. Dans le second cas, les symboles j, k et w ne se produisent jamais; notons qu'on ne tient pas compte des distinctions de signes diacritiques.

7. Dans le corpus de *contrôle* - c'est-à-dire incluant la segmentation en mots - nous avons remplacé les symboles non alphabétiques par des espaces avant de réduire les séquences d'espaces à un espace unique; notons que le texte anglais paraît plus sérieusement corrompu par cette manipulation, essentiellement parce qu'elle entraîne des erreurs relativement fréquentes liées à la suppression abusive de certains apostrophe et tirets (voir paragraphe 1.1).

d'ordre k pour un alphabet de taille $m$doivent être estimés à partir d'un corpus de longueur au moins égale à m^{k+1}, on peut facilement déduire que ce ne sont pas moins de 457'000 symboles qui sont nécessaires en anglais pour l'entraînement à l'ordre 3, contre près de 280'000 pour le latin. Nous avons donc supprimé une partie (la fin) de chaque texte en ne conservant qu'une marge symbolique avec ces chiffres; les corpus définitifs, que nous désignerons comme AUSTEN et OVIDE, mesurent respectivement 618'696 et 352'419 symboles avec les espaces, et 500'024 et 300'006 après leur suppression.

Nous sommes finalement revenus sur la décision de s'arrêter à l'ordre 3, et avons constaté que le passage à l'ordre supérieur permettait une amélioration remarquable des résultats, malgré le diagnostic pessimiste de la borne de fiabilité de l'estimation. Ces résultats sont donc inclus dans la version définitive de ce travail - précédés de la mention explicite de leur statut particulier.

4.3 Evaluation

Dans le contexte *bruité* que nous avons évoqué à plusieurs reprises, spécifier une procédure objective d'évaluation des résultats n'est pas la moindre des difficultés à laquelle nous devons faire face[8]. L'option de segmentation (en *mots*) que nous avons retenue résulte essentiellement de cette contrainte. A cet égard, l'intérêt évident des corpus orthographiques est de permettre l'utilisation des espaces pour approximer la position des frontières de mots.

Si l'on accepte cette simplification, nous disposons d'un moyen direct d'évaluer la qualité d'une segmentation, en utilisant les mesures intuitives de la théorie de la détection du signal. Connaissant la véritable segmentation du corpus, nous pouvons classer chaque décision du système dans l'une des catégories figurant sur le tableau 4.1 ci-dessous.

En sommant sur les colonnes, on obtient le nombre de frontières réelles *vp* + *fn* et celui de «non-frontières» réelles *fp* + *vn*. Sur les lignes, c'est le nombre de frontières détectées *vp* + *fp* et non détectée v*n*+ *fn* que l'on calcule. Si enfin la sommation est effectuée sur les diagonales, on obtient le nombre de décisions correctes *vp* + *vn* et celui d'erreurs *fn* + *fp*. Le taux d'erreur:

$$\text{error} := \frac{fp + fn}{vp + fp + fn + vn} = \frac{fp + fn}{n - 1} \tag{4.1}$$

8. d'autant plus qu'évaluation et application de la procédure sont étroitement liés, puisque nous n'avons pas, pour le moment, de solution plausible pour la détermination automatique d'une valeur de seuil

		Réalité	
		Frontière	¬ **frontière**
Détection	**frontière**	*vrai positif* (*vp*): correct	*faux positif* (*fp*): fausse alarme
	¬ **frontière**	*faux négatif* (*fn*): manqué	*vrai négatif* (*vn*): correct

TAB. 4.1 – *Evaluation en terme de détection de signal*

mesure la proportion de décisions fausses sur le total des décisions (n dénote la taille du corpus). C'est la mesure que nous chercherons à minimiser en faisant varier le seuil $T(\sigma)$ [9]. Mais cela n'est pas toujours une bonne solution, puisque dans certains cas, cette stratégie aboutit à la sélection d'un seuil maximal, tel que le critère $d(\sigma,i)$ lui est systématiquement inférieur [10]. Dans le cas d'OVIDE, par exemple, on obtient un taux d'erreur de 0.175 en n'insérant aucune frontière, c'est-à-dire que le modèle «prend la bonne décision» plus de quatre fois sur cinq; en général, pour l'application considérée, le taux d'erreurs favorise une segmentation *conservatrice*, en ce sens qu'elle tend à sélectionner un seuil auquel échappent de nombreuses frontières réelles, mais qui produit également peu de fausses alarmes.

Dans le cas problématique où le critère du taux d'erreur aboutit à un modèle *silencieux*, nous recourrons à d'autres mesures, plus détaillées; la probabilité que le système détecte correctement une frontière existante (*recall*) et celle qu'il détecte une frontière à tort (*fallout*) sont données par:

$$\text{recall} = \frac{vp}{vp + fn} \quad \text{fallout} = \frac{fp}{fp + vn} \tag{4.2}$$

Dans l'idéal, on souhaite que le système détecte toutes les frontières existantes et aucune autre (recall = 1, fallout = 0). Mais la situation est souvent plus nuancée: le fallout n'est généralement pas nul lorsque le recall est égal à l'unité (inversement, si le fallout est nul, le recall n'est généralement pas égal à l'unité), ce qui signifie que le système ne peut détecter toutes les frontières existantes sans induire de fausses alarmes. Il s'agit alors de trouver un compromis entre la capacité du système de rester silencieux en l'absence de signal, et celle de détecter sa présence.

9. Voir partie 2.

10. ou supérieur, dans le cas d'une approche basée sur la dissimilarité (voir partie précédente)

Le seuil $T(\sigma)$ joue le rôle d'arbitre dans ce compromis. En le faisant varier de $-\infty$ à $+\infty$, on peut définir toutes les paires recall - fallout possibles, et les représenter comme une série de points sur un plan. Le résultat est appelé courbe ROC (pour *receiver operating characteristic*) et constitue une représentation graphique du comportement perceptif du système[11]. Sur la figure 4.1 (page suivante), où chaque courbe correspond à une variable de décision, on peut voir que la variable D est la plus efficace: en variant le seuil, il est possible d'augmenter le recall jusqu'à 80% sans augmentation significative du fallout - inversement, il suffit de réduire le recall à 80% pour que le fallout devienne négligeable.

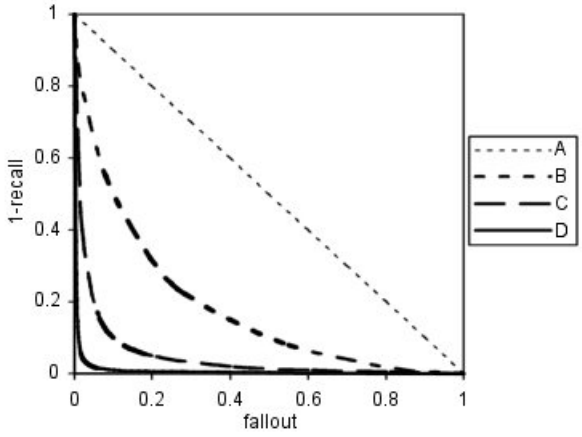

FIG. 4.1 – *courbes ROC*

Revenant à notre problème de mutisme occasionnel du système pour minimiser le taux d'erreur, nous proposons de sélectionner dans ce cas le seuil $T(\sigma)$ tel qu'il minimise la différence (en valeur absolue) | fallout - (1-recall) |. Ceci revient à accorder la même importance aux deux types d'erreurs, hypothèse somme toute raisonnable si l'on ne sait rien de plus sur les phénomènes considérés.

Notons encore que lorsque la procédure de sélection aboutit finalement à un *choix* de valeurs, nous tranchons en choisissant celle qui minimise le fallout.

11. Il existe plusieurs variantes de cette courbe; dans ce mémoire, nous rapportons le fallout (en abscisse) au *complément* du recall, soit la probabilité de manquer une frontière existante.

Cette approche[12] reflète notre intuition qu'il est plus prudent de s'en tenir aux indices *fiables*, quitte à manquer une partie des frontières ciblées. Nous reviendrons sur ce point dans notre conclusion, et songerons à une façon d'envisager la récupération *ultérieure* des frontières manquées en généralisant les informations déjà acquises.

4.4 Discussion des résultats

La remarque la plus générale que nous puissions faire quant aux résultats de notre expérimentation est que la qualité de la segmentation[13] (telle qu'évaluée par le taux d'erreur) augmente toujours avec l'ordre du modèle; c'est ce que confirment tous les graphiques de l'annexe B.2 (p.66 et suivantes). Ceci va dans le sens des constatations de Harris[14], et traduit directement le gain d'information lié au passage à un ordre supérieur - dans les limites de représentativité des indicateurs.

Le taux d'erreur le plus faible est atteint à l'ordre 4 par la divergence: 7.5% pour le corpus AUSTEN, et 10.5% pour le corpus OVIDE. Cette prééminence de la divergence sur les autres indicateurs vérifie sans surprise notre hypothèse quant à la possibilité d'améliorer les performances du système en tenant compte de la *vraie* distribution initiale - estimée sur les corpus avant la suppression des séparateurs. Toutefois, nous verrons plus loin que ce n'est pas forcément le cas aux ordres inférieurs. La segmentation résultante[15] est assez *conservatrice*, en ce sens qu'elle manque un certain nombre de frontières (22.9% et 46% respectivement), mais ne produit que peu de fausses alarmes (2.6% et 3%). Celles-ci sont principalement le fait d'*homonymies* partielles et de phénomènes *d'affixation* (angl. *h_and_some*, *excellen_**t** **wo**man*, *s_he*, *in_distinct*, *comfort_able*, *year_s*, lat. *ere_ctos*, *nullo s_ponte*, *me_a*, *per_missit*, *utram_que*).

Sur les figures B.3 et B.4 (pp. 66-66), on peut voir que l'utilisation de la divergence impose une limite supérieure au fallout et inférieure à la proportion de frontières manquées (1-recall): par exemple, pour le corpus OVIDE à l'ordre 4, on ne peut pas faire plus de 55.1% de fausses alarmes ni manquer moins de 22% des frontières réelles. Ces limites proviennent de la définition de la divergence, qui attribue une valeur infinie (donc ne pouvant passer en dessous d'aucun seuil) aux contextes où peut se produire l'un au moins des symboles interdits en position initiale (voir paragraphe 3.1).

12. de même que le fait de choisir le taux d'erreur comme critère principal
13. Nous ne préciserons plus dans la suite que c'est de segmentation en *mots* qu'il s'agit ici.
14. Voir note 6, p. xvi et alinéa 2.2.1.
15. extraits o4-K-1.38 p. 70 et o4-K-1.14 p. 74

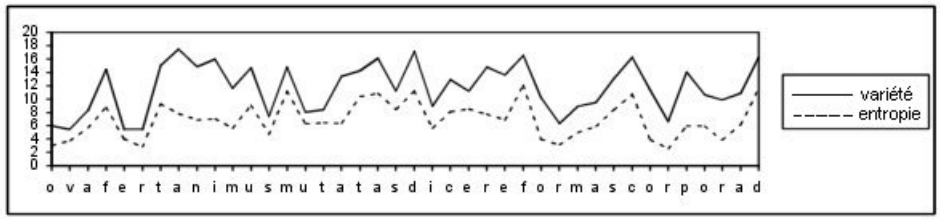

FIG. 4.2 – *Variété et entropie conditionnelle (ordre 3), corpus OVIDE*

Une autre observation de portée très générale est que la variété fait systématiquement moins bien que les indicateurs tenant compte des fréquences[16]. L'examen du détail de la courbe de la variété par rapport à celle de l'entropie (voir figure 4.1, p. 38[17]), montre que la valeur de l'entropie est toujours inférieure ou égale à celle de la variété (par définition), sans que cela altère fondamentalement les rapports entre valeurs successives pour autant. Mais on constate également que l'écart n'est pas constant; en particulier, les pics de variétés observés entre *an* et *imus*, *dice-* et *-re* ou *cor* et *pora* semblent être *corrigés* sur la courbe de l'entropie, ce qui signifie qu'à ces endroits, si de nombreux successeurs (ou prédécesseurs) sont effectivement attestés, leur distribution est trop écartée de l'uniformité pour qu'on les fasse suivre (ou précéder) d'une frontière de mot.

Un exemple intéressant de ce phénomène se présente lorsqu'un suffixe flexionnel, utilisé particulièrement souvent (et donc au détriment des autres suffixes du même paradigme), fait chuter l'entropie par rapport à la variété, évitant ainsi de produire une fausse alarme. Dans ce cas, prendre les fréquences en considération aboutit indirectement à améliorer la discrimination entre les degrés d'*indépendance* mentionnés par Harris (1955a, pp.60-1). Notons que dans son approche, dont le but est la segmentation en *morphes*, l'emploi de l'entropie engendrerait un accroissement du taux d'erreur, puisque les problèmes d'indépendance réduite (alternance, affixation, etc.) nuisent généralement à la détection des frontières de morphes, et que les indicateurs tenant compte des fréquences renforcent cet effet.

16. en particulier pour les ordres faibles

17. en l'occurrence, ce n'est pas l'entropie conditionnelle qui est représentée, mais la quantité $2^{\vec{h}(w)}$; celle-ci croît et décroît comme l'entropie, mais varie entre 1 et m (la taille de l'alphabet A) plutôt qu'entre 0 et $\log m$.

En ce qui concerne l'hypothèse portant sur la plus grande généralité de l'information mutuelle par rapport à l'entropie, on remarque dès l'abord que la première aboutit à une meilleure segmentation que la seconde pour le latin et inversement pour l'anglais (à l'exception de l'ordre 2, voir figure 4.3 (page suivante). En latin, pour les ordres inférieurs à 4, l'information mutuelle fait même mieux que l'approche *supervisée* basée sur la divergence.

	Variété		Entropie		Information mutuelle		Divergence	
Ordre	AUSTEN	OVIDE	AUSTEN	OVIDE	AUSTEN	OVIDE	AUSTEN	OVIDE
2	23.4%	(36.9%)	18.1%	(26.8%)	16.8%	16.1%	16.4%	17.9%
3	12.8%	(23%)	10.8%	15.5%	11.2%	13.8%	10.5%	14.2%
4	10.1%	14.6%	8.4%	13.3%	9.4%	13.1%	7.4%	10.5%

FIG. 4.3 – *Taux d'erreur* [18]

Là où l'entropie prédomine, c'est-à-dire en anglais et pour les ordres supérieurs à 2, elle ne prend pas une avance décisive sur l'information mutuelle (0.4% à l'ordre 3 et 1% à l'ordre 4). En contraste, l'information mutuelle est souveraine à l'ordre 2, en particulier pour le corpus OVIDE: dans ce cas, la variété et l'entropie échouent à induire une diminution du taux d'erreur (voir figure B.2, p. 65). En considérant la moyenne du taux d'erreur sur les deux corpus pour chaque indicateur et chaque ordre, on voit que - selon nos critères de sélection - l'information mutuelle donne les meilleurs résultats aux ordres 2 (16.5% contre 22.5% pour l'entropie) et 3 (12.5% contre 13.2%), tandis que l'entropie prévaut à l'ordre 4 (10.9% contre 11.3% pour l'information mutuelle). En généralisant à tous les ordres, on obtient un taux d'erreur moyen global de 13.4% pour l'information mutuelle et 15.5% pour l'entropie [18], ce qui est en définitive notre plus fort argument pour accepter l'hypothèse que la première a une portée plus générale que la seconde, du moins pour les corpus considérés; cette tendance semble être particulièrement nette aux ordres faibles, pour s'amenuiser en passant aux ordres supérieurs [19].

18. Ce dernier pourcentage est biaisé par le cas du corpus OVIDE à l'ordre 2, où l'entropie obtient le moindre taux d'erreur (17.5%) en n'insérant aucune frontière, et où l'on a donc appliqué le critère de la moindre différence absolue | fallout - (1-recall) |; notons que même en remplaçant le taux obtenu dans ce cas par le moindre taux, on obtient un taux moyen supérieur de 0.5% à celui de l'information mutuelle.

19. Il serait d'intéressant de savoir ce qu'il advient de cette diminution de l'écart au-delà des ordres testés ici.

Dans l'ensemble, l'information mutuelle aboutit à une segmentation plus conservatrice que l'entropie, évitant les fausses alarmes pour favoriser le recall. Nous reviendrons dans la partie suivante sur une façon d'exploiter cette circonstance.

Chapitre 5

Conclusions et perspectives

Arrivés au terme de cette exploration, nous sommes désormais en mesure de dresser le bilan du chemin parcouru, avant de passer en revue les aspects du modèle qui restent à développer, et les extensions suggérées par nos résultats.

5.1 Résumé

Nous avons d'abord cherché à donner une introduction informelle au problème de la segmentation morphologique, en évoquant notamment sa pertinence dans le cadre de la linguistique descriptive, des théories de l'acquisition et du traitement automatique des langues naturelles. Puis nous avons introduit les contraintes dans le cadre desquelles notre algorithme de segmentation doit pouvoir opérer, portant en particulier sur le caractère segmental, continu, linéairement séparable et manifeste du signe morphologique; nous avons insisté sur les limites absolues qu'elles imposent à la segmentation. Ensuite, nous avons ouvert une première discussion du problème spécifique qui motive cette recherche; nous avons vu ainsi que la méthode du nombre de successeurs décrite par Harris (1955b) - et dont de nombreux travaux ultérieurs se sont inspirés - n'est en fait qu'un cas particulier d'une approche basée sur la similarité entre la distribution (*conditionnelle*) des successeurs d'un contexte et la distribution des symboles en début d'énoncé, qui sert d'approximation de la distribution (*initiale*) des symboles en début de mot; l'utilisation d'une mesure de diversité revient à faire l'hypothèse que la diversité de la distribution initiale (généralement inconnue) est maximale, et que ce critère suffit à évaluer la similarité en question. Nous avons conclu notre introduction en explicitant les deux hypothèses qui devaient

orienter l'ensemble des considérations subséquentes:

a. qu'il devrait être possible d'améliorer les performances d'un algorithme de segmentation si l'on connaît la vraie distribution initiale;

b. que la distribution (*non conditionnelle*) des symboles *hors contexte* pourrait constituer une meilleure approximation de la distribution initiale en l'absence d'autre information *a priori*.

Notre deuxième partie fut essentiellement consacrée à la formulation rigoureuse d'un algorithme générique de segmentation par seuillage, dont nous avons d'abord montré la relation avec la méthode du nombre de successeurs avant de définir plusieurs contraintes sur le conditionnement de la variable testée. Nous avons ainsi évoqué le *bruit* induit par les hypothèses (dites *markoviennes*) de *stationnarité* et de *conditionnement limité* - mal nécessaire pour réduire le nombre de paramètres à estimer, donc la taille du corpus d'entraînement nécessaire (et le temps de calcul, accessoirement). Nous avons également mentionné la simplification adoptée dans ce travail et consistant à utiliser un seuil constant pour l'ensemble de la séquence, plutôt qu'un seuil adapté en fonction du contexte immédiat de la frontière testée. Sur cette base, il fut possible de dériver deux premiers critères markoviens pour la segmentation: la variété et l'entropie conditionnelle (à un k-gramme). Enfin, nous avons cherché à intégrer dans le formalisme la notion de *précédence* [1], et avons proposé de combiner indicateurs transitionnels et précédentiels de façon simple et intuitive sous la forme d'une moyenne (éventuellement pondérée).

C'est dans la partie 3 que nous avons formalisé le développement spécifique que nous proposons, à savoir d'expliciter le caractère différentiel de la méthode de Harris en utilisant une mesure de dissimilarité, et d'utiliser comme approximation de la distribution initiale la distribution non conditionnelle plutôt qu'une distribution uniforme (comme dans le cas de l'entropie). Nous avons vu ainsi que ce changement de point de vue, apparemment radical, se traduit formellement par la dérivation très directe de la divergence de Kullback-Leibler à partir de l'entropie conditionnelle. Puis nous avons tenté de justifier l'approximation de la distribution initiale par la distribution non conditionnelle en montrant qu'elle revient à fonder la segmentation sur la dépendance entre le contexte et ses successeurs (et à utiliser l'information mutuelle au lieu de l'entropie).

Dans la partie 4, nous avons tenté d'évaluer la portée de nos hypothèses et la pertinence des développements proposés en les soumettant à l'épreuve des

1. Dont l'utilisation est suggérée par Harris dès son article de 1955, sous la forme du nombre de *prédécesseurs*.

données - pour une tâche de segmentation en mots de corpus orthographiques. Nous avons d'abord discuté la nature des corpus sélectionnés et la norme de codage (simpliste) adoptée, et présenté les mesures retenues pour l'évaluation des performances du modèle: taux d'erreur, fallout et recall. Nous avons convenu de sélectionner la valeur du seuil qui minimise le taux d'erreur, à moins que ce choix n'aboutisse au mutisme du système, auquel cas nous cherchons à minimiser la différence absolue | fallout - (1-recall) | ; s'il subsiste un choix après ce premier filtrage, nous sélectionnons le seuil qui minimise le fallout. En examinant les résultats au moyen de ces outils, nous avons pu constater que notre première hypothèse est largement vérifiée pour les corpus testés, c'est-à-dire que (la minimisation de) la divergence entre la distribution conditionnelle et la *vraie* distribution initiale est un meilleur indice pour la segmentation que (la maximisation de) la diversité ou l'incertitude. En revanche, la divergence fait moins bien que la méthode basée sur l'indépendance pour les ordres inférieur à 4; en outre, cette seconde approche ne donne pas des résultats beaucoup moins bons que l'entropie sur le corpus anglais, et fait mieux que tous les autres à l'ordre 2. Nous y voyons des raisons convaincantes de considérer que la dépendance est un critère plus général que la diversité (ou l'incertitude), et donc d'y recourir lorsqu'on n'a pas de raisons de penser que l'on se trouve dans le cas particulier (quoique fréquent) d'une langue où la diversité suffit à rendre compte de la similarité entre distribution conditionnelle et initiale.

5.2 Problèmes en suspens, développements possibles

Plusieurs aspects du formalisme proposé dans ce mémoire sont en attente d'un traitement plus systématique. C'est le cas en particulier de la question du seuil, dont nous avons déjà relevé qu'elle est l'une des plus régulièrement escamotées dans ce contexte. L'indétermination porte bien sûr sur sa forme: le seuillage doit-il s'appliquer aux valeurs élevées ou aux pics? ou encore, aux pics élevés, aux coudes, etc.? Mais elle concerne plus encore la question épineuse de la valeur effective du seuil: comment choisir une bonne valeur de seuil lorsqu'on ne dispose pas, comme c'était notre cas dans la partie 4, d'un moyen d'évaluer les résultats? De toute évidence, ces problèmes sont d'une importance cruciale pour notre propos, et nous ne pouvons que regretter de n'avoir pas le temps de les examiner plus spécifiquement dans ce travail.

Il serait également intéressant de creuser la question du conditionnement de la variable. Nous avons mentionné dans l'alinéa 2.2.2.3 une possibilité d'extension (le cas des quantités précédentielles), mais il serait possible également

d'augmenter l'ordre de l'unité *prédite*. Par exemple, à partir d'une distribution de 4-grammes, on peut estimer la probabilité qu'un trigramme donné soit suivi ou précédé par un symbole donné, mais aussi celle qu'un bigramme soit suivi ou précédé par un autre - et rien n'indique que la seconde estimation ne soit pas mieux adaptée que la première pour la segmentation.

L'un des attraits d'un algorithme de segmentation basé sur des indicateurs conditionnés sur une distribution de k-grammes (comme les versions proposées dans ce travail de la variété et l'entropie conditionnelle, et de l'information mutuelle) est la facilité avec laquelle on peut rendre compte de l'évolution du modèle dans le temps. On pourrait aisément imaginer une méthodologie permettant de décrire l'apprentissage de ce système *à mesure qu'il traite les données*. Dans le même ordre d'idée, nous n'avons pas du tout cherché à déterminer dans quelle mesure les paramètres estimés sur un corpus peuvent être utilisés pour en segmenter un autre.

Surtout, nos résultats suggèrent qu'il est difficile d'obtenir des performances réellement satisfaisantes par un passage unique, sans nuance, sur le corpus. De façon générale, les paramètres qui minimisent le taux d'erreur de notre système tendent à sélectionner des valeurs de seuils telles qu'elles échouent à détecter de nombreuses frontières existantes, mais qu'elles n'induisent que peu de fausses alarmes. C'est cette circonstance qui nous conduit à postuler la nécessité d'une approche *itérative* de la segmentation. Supposons que l'on utilise un algorithme basé sur la divergence, donc sur deux modèles explicites, et qu'on choisisse comme première approximation de la distribution initiale la distribution non conditionnelle. Après une première itération, nous obtenons une première segmentation du corpus, à partir de laquelle il est possible réestimer la distribution initiale. Une fois le modèle mis à jour, on peut pratiquer une nouvelle segmentation - possiblement plus correcte [2] - et ainsi de suite jusqu'à ce que la distribution initiale soit stabilisée.

Cet algorithme hypothétique est l'une des extensions que nous souhaitons pouvoir expérimenter ultérieurement. Il nous semble tirer le meilleur parti de la notion d'apprentissage par *généralisation*, dont nous avons vu plus haut qu'elle est le fondement de la méthode de Harris. Pour l'heure, c'est d'une conclusion qu'il doit s'agir ici, et à ce titre nous pensons pouvoir dire - au risque d'une dernière répétition - que nous avons montré:

a. que la méthode de nombre de successeurs est un cas particulier d'une approche plus générale et basée sur la notion de (dis-)similarité entre distribution conditionnelle et initiale;

2. C'est sur ce point que repose tout le raisonnement; cette approche n'a pas de sens si la distribution initiale ne converge pas vers un optimum pour la segmentation.

b. que la distribution non conditionnelle est une alternative intéressante à la distribution uniforme (pour approximer la distribution initiale), et vraisemblablement plus générale;

c. qu'une fois de plus, il n'a pas vraiment été tenu compte des premiers travaux en la matière lors des développements ultérieurs, et - nous l'espérons - que le genre de rétrospection que nous avons pratiqué dans ces pages peut faire émerger des hypothèses et présupposés noyés dans la chaîne des paraphrases bibliographiques.

Annexe A

Introduction à l'analyse des séries temporelles catégorielles

Dans cette partie, nous exposerons les éléments du formalisme de l'*analyse des séries temporelles catégorielles* (ASTC[1]) qui fondent l'approche entropique de la segmentation morphologique. La présentation sera axée essentiellement sur les rapports existant entre le modèle des chaînes de Markov (Markov 1916) et la Théorie de l'Information (Shannon 1951). Nous postulons que les notions de base de la statistique classique (en particulier le calcul des probabilités) sont familières au lecteur - et recommandons au néophyte une myriade de bonnes introductions, parmi lesquelles celle de Bavaud (1998) présente l'intérêt de développer certains thèmes spécifiques de l'ASTC.

A.1 Modélisation probabiliste du texte

Dans le cadre de l'ASTC, on appelle *texte*[2] (de longueur n) tout échantillon ordonné $\sigma := s_1^n := s_1...s_n$, constitué par la concaténation de n états successifs d'une variable catégorielle S codée sur un *alphabet* $A := \{a_1,...,a_m\}$ de m modalités ou *symboles*[3]. Selon cette définition, le caractère *textuel* d'un échantillon

1. L'usage de l'acronyme dénote moins le figement du terme que sa lourdeur rédactionnelle; notons toutefois que l'ASTC correspond plus ou moins à la branche anglo-saxonne des recherches en statistique textuelle - par opposition aux travaux français axés essentiellement sur l'analyse des correspondances (voir Bavaud 2000, p.263).

2. ou *séquence*, ou encore *série temporelle catégorielle*

3. Cette définition recouvre des types de données que l'usage ordinaire du terme tend à exclure, comme le résultat d'une série de lancers à pile ou face, par exemple.

relève pour une part de son déroulement sur un axe temporel discret, et pour l'autre de sa nature catégorielle.

Pour un texte σ donné, l'ASTC vise essentiellement à répondre à deux questions:

a. Comment *décrire*, de façon simple et explicative, le processus ayant généré les données observées?

b. Dans quelle mesure le modèle textuel ainsi conçu permet-il de *prédire* les états futurs s_{n+1},s_{n+2},...?

La notion de modèle textuel est fondamentale dans le cadre de l'ASTC; qu'on l'envisage en tant que résultat de la description d'un texte ou comme condition de sa prédiction, un modèle textuel consiste en un corps d'assertions portant sur le déroulement d'un processus textuel, conditionnées par des informations contextuelles plus ou moins détaillées. Si ces assertions produisent une prédiction univoque, elles équivalent à des expressions de la forme $s_{t+1} = f(t,c^t)$, où $c^t = s_1 \ldots s_n$ représente la totalité des symboles précédant le symbole à prédire [4] et $f(t,c^t)$ une fonction associant à tout contexte c^t au temps t un symbole prédit unique. Un modèle ainsi formulé est dit *déterministe* et s'avère particulièrement vulnérable en ce sens qu'il suffit d'une contre-observation pour le réfuter. Pour cette raison [5], on recourt souvent à une expression *probabiliste* du modèle; dans ce cas, les assertions sont notées $p_i = p(s_{t+1} = a_i \mid t,c^t)$, et caractérisent la probabilité conditionnelle d'apparition du symbole $a_i \in A$ étant donné le contexte c^t au temps t. Pour t et c^t fixés, elles sont liées entre elles par $\sum_{i=1}^{m} p_i = 1$ et constituent une distribution $p_1, \ldots, p_m$ sur l'alphabet A. En fait, les assertions déterministes définies précédemment correspondent à un type extrême de distributions, soit lorsque toutes les probabilités sont concentrées dans un symbole unique.

Pour simplifier le modèle probabiliste introduit à l'instant, on formule généralement l'*hypothèse de stationnarité* qui revient à considérer que la probabilité d'un symbole $a \in A$ ne dépend pas explicitement de l'instant t où s'effectue la prédiction, mais uniquement du contexte c^t [6]:

4. Pour simplifier la présentation, et parce que nous plaçons dans le cadre d'une application de prédiction, nous limitons ici le contexte aux symboles passés (voir paragraphe 2.1.4 pour une approche alternative); de même, nous limitons l'objet prédit au seul symbole suivant le contexte, mais le formalisme est généralisable à des prédictions plus étendues.

5. et pour d'autres (voir Bavaud 1998, pp.87-8, pour une discussion plus approfondie)

6. On peut s'attendre à ce que la qualité de la prédiction soit diminuée, mais l'hypothèse s'avère indispensable en pratique pour réduire la dimensionnalité du problème, notamment

$$p(a \mid c) := p(s_{t+1} = a \mid t,c^t) \approx p(s_{t+T+1} = a \mid t+T,c^{t+T}) \qquad T = 1,2,3,... \tag{A.1}$$

Une seconde simplification usuelle, l'*hypothèse de conditionnement limité*, postule que la connaissance de la totalité du contexte c n'influence pas plus la probabilité d'un symbole que la connaissance des k derniers symboles de c[7]. Ainsi, si l'on note $w \in A^k$ le dernier *k-gramme* ou groupe de k symboles de c (A^k) représente l'ensemble des k-grammes possibles), on a:

$$p(a \mid c) \approx p(a \mid w) \tag{A.2}$$

Dans ces conditions, on définit la *probabilité de transition de w vers a* comme:

$$p(w \to a) := p(a \mid w) = \frac{p(wa)}{p(w)} \tag{A.3}$$

où wa désigne le résultat de la concaténation de w et a. En spécifiant $p(w \to a)$ pour tout $w \in A^k$ et $a \in A$, on peut construire une matrice de transition $P(m^k \times m)$ de composantes $P_{wa} := p(w \to a)$, qui définit une *chaîne de Markov* d'ordre k[8]. Dans le cas extrême où k est nul, on aboutit à modèle d'ordre 0 où la probabilité d'un symbole est indépendante du contexte de la prédiction:

$$p(\varepsilon \to a) = p(a) \tag{A.4}$$

où ε dénote la *chaîne vide* qui constitue à elle seule l'ensemble des 0-grammes.

A.2 Information, entropie et chaînes de Markov

L'ASTC fait un usage immodéré des concepts de la Théorie de l'Information. Une exposition exhaustive de ce formalisme dépasse largement la portée de ce mémoire - et à plus forte raison, de cette annexe. Nous nous contenterons donc de considérer la transposition des concepts informationnels dans le cadre

lorsqu'il s'agit d'estimer les paramètres du modèle (voir annexe A.3); dans la suite, elle sera systématiquement sous-entendue, de même qu'on évitera d'expliciter répétitivement le caractère *stationnaire* des probabilités résultantes.

7. Ici encore, il s'agit d'une réduction d'information consentie au profit d'un allégement du modèle.

8. Dans la suite, nous parlerons quelquefois de distribution *conditionnelle étant donné w* (ou distribution *des successeurs de w*) pour désigner une ligne particulière de la matrice de transition correspondante:
$p^w := p(w \to a_1),...,p(w \to a_m)$.

d'un modèle textuel probabiliste. Le lecteur soucieux de comprendre les rouages fondamentaux de la Théorie de l'Information est invité à consulter un ouvrage de référence comme Cover & Thomas (1991)[9].

Pour commencer, considérons la question suivante: étant donné une distribution de probabilités sur l'alphabet $A := \{a_1,...,a_m\}$, comment qualifier l'*information* $i(a)$ produite par l'apparition d'un symbole $a \in A$ donné? On comprend intuitivement que $i(a)$ devrait être nulle ssi la probabilité $p(a)$ du symbole correspondant est égale à 1 (puisque l'on *sait* qu'il va se produire), et inversement qu'elle devrait tendre vers ∞ pour $p(a)$ tendant vers 0. On souhaiterait en outre que l'information liée à l'observation de deux symboles successifs a ,$a' \in A$ soit égale à la somme de l'information associée à chaque symbole individuellement si et seulement s'ils sont indépendants:

$$i(aa') = i(a) + i(a') \Leftrightarrow p(aa') = p(a)p(a') \tag{A.5}$$

Cette propriété évoque immédiatement celles du logarithme, et conduit à la définition suivante, initialement proposée par Hartley (1928):

$$i(a) = \log \frac{1}{p(a)} = -\log p(a) \tag{A.6}$$

pour tout $a \in A \ / \ p(a) \neq 0$[10], et où l'on utilise le logarithme binaire[11]. L'*information* ainsi formulée vérifie (A.5), et l'on a toujours:

$$0 \leq i(a) \leq \infty \ \text{ et } \ i(a) = \log m \Leftrightarrow p(a) = \frac{1}{m} \tag{A.7}$$

L'information se généralise naturellement au cas des k-grammes $w \in A^k$: $i(w) := -\log p(w)$. La dérivation de l'information *conditionnelle* $i(w \rightarrow a)$[12] produite par l'apparition d'un symbole a donné après un contexte w donné fait intervenir la propriété (A.5):

9. On trouvera également de nombreux éléments dans Welsh (1988) et un résumé intéressant dans Manning & Schütze (1999, pp.60-80); pour les aspects spécifiquement textuels, nous renvoyons à nouveau à Bavaud (1998).

10. Dans la suite, sauf mention explicite, nous admettrons que toutes les opérations décrites concernent les probabilités non-nulles.

11. C'est là un usage courant en Théorie de l'Information; changer de base revient à exprimer l'information dans une autre unité: *bits* (base 2), *nats* (base e) ou *digits* (base 10).

12. La notation que nous utilisons est peu conventionnelle; nous l'adoptons pour mettre en relief le rapport direct entre probabilité et information correspondantes.

$$i(w \to a) = -\log p(w \to a) = -\log \frac{p(wa)}{p(w)}$$
$$= \quad -\log p(wa) \quad + \quad \log p(w) \quad = \quad i(wa) \quad - \quad i(w) \quad \text{(A.8)}$$

Comme on le voit, $i(w \to a)$ s'interprète comme la différence entre l'information associée au k+1-gramme wa et celle associée au seul k-gramme w. Elle est nulle ssi $p(wa) = p(w)$, c'est-à-dire si $p(w \to a) = 1$.

L'information *mutuelle ponctuelle* $I(w \to a)$ [13] entre un k+1-gramme w et son successeur a est donnée par:

$$I(w \to a) = -\log \frac{p(wa)}{p(w)p(a)} = -\log \frac{p(w \to a)}{p(a)} \quad \text{(A.9)}$$

$I(w \to a)$ varie entre $-\infty$ et $+\infty$, et s'annule ssi $p(wa) = p(w)p(a)$; en ce sens, elle constitue une mesure de la dépendance entre w et a. Elle peut aussi être conçue comme la différence entre l'information associée à une transition $w \to a$ et celle associée au seul symbole a [14], ou encore comme la différence entre l'information associée au k+1-gramme wa et la somme de celles associées au k-gramme w et au symbole a indépendamment, en vertu de:

$$I(w \to a) = i(wa) - (i(w) + i(a)) \quad \text{(A.10)}$$

La quantité $i(a)$ définie par (A.6) caractérise l'information associée à un symbole donné. Sur l'ensemble d'une distribution - contrairement aux probabilités correspondantes, la somme des $i(a)$ n'est pas contrainte de valoir 1. Il en résulte que l'espérance mathématique de l'information associée à un symbole $a \in A$ tiré au hasard $E_a(i(a)) = \sum_a p(a)i(a)$ n'est généralement pas la même pour deux distributions différentes (à l'inverse de $E_a(p(a)) = \frac{1}{m}$ par définition) [15]. Par exemple, considérons l'information produite en moyenne par le lancer d'une pièce équilibrée et celui d'une pièce pour laquelle la probabilité de l'une des issues (disons 'pile') vaut 0.8: dans le premier cas, l'information produite par l'une et l'autre des issues (équiprobables) vaut 1 bit d'après (A.7), d'où

13. Il convient de distinguer cette notation de la précédente (A.8).

14. A ce sujet, notons que l'information mutuelle ponctuelle est *symétrique* en ce sens que l'on a toujours $I(w \to a) = i(w \to a) - i(a) = i(w \leftarrow a) - i(w) = I(w \leftarrow a)$ (l'utilisation de la flèche inversée dénote le caractère *précédentiel* de la quantité correspondante, au sens défini dans l'alinéa 2.1.4); en revanche, on a généralement $I(w \to a) \neq I(a \leftarrow w)$.

15. L'indice a sous la sommation est un raccourci informel pour "$a \in A : p(a) \neq 0$" (voir note 10, p. 52).

$E_a(i(a)) = 1$ bit également; dans le second, on a $i(pile) = -\log 0.8 = 0.32$ bits et $i(face) = -\log 0.2 = 2.32$ bits, d'où $E_a(i(a)) = (0.8 \cdot 0.32) + (0.2 \cdot 2.32) = 0.72$ bits. En moyenne, on obtient moins d'information en lançant la pièce biaisée, et l'on n'en obtiendrait pas du tout si elle ne produisait que 'pile'.

Shannon (1951) a défini l'*entropie* $H(S)$ associée à la distribution $p := p(a_1), \ldots, p(a_m)$ des symboles de S comme l'espérance mathématique de $i(a)$:

$$H(S) := H(p) := E_a(i(a)) = \sum_a p(a) i(a) = -\sum_a p(a) \log p(a) \tag{A.11}$$

Le mérite lui revient d'avoir démontré que l'entropie présente un ensemble de propriétés *souhaitables* [16] pour une mesure de l'*incertitude* ou du *manque d'information associé à une distribution* [17], dont la plus importante est incontestablement la suivante: elle est maximale et égale à $\log m$ bits pour une distribution uniforme, et minimale et nulle pour une distribution déterministe:

$$\begin{aligned} 0 \leq H(p) &\leq \log m \\ H(p) = 0 &\Leftrightarrow p \text{ déterministe} \\ H(p) = \log m &\Leftrightarrow p \text{ uniforme} \end{aligned} \tag{A.12}$$

Dans le cadre de l'ASTC, on généralise (A.11) aux distributions p^k de k-grammes $w \in A^k$ en définissant l'entropie *d'ordre* k (pour $k \geq 1$) comme:

$$H_k := H(p^k) = E_w(i(w)) = -\sum_w p(w) \log p(w) \tag{A.13}$$

Il découle des propriétés de l'entropie que l'on a toujours $H_{k+1} \geq H_k$, avec égalité ssi toutes les distributions conditionnelles étant donné un k-gramme [18] $w \in A^k$ sont déterministes:

$$H_{k+1} = H_k \Leftrightarrow (\forall\, w \in A^k) H(p^w) = 0 \tag{A.14}$$

L'entropie *conditionnelle* étant donné w est définie comme:

$$h(w) := H(p^w) = E_a(i(w \to a)) = -\sum_a p(w \to a) \log(w \to a) \tag{A.15}$$

16. On trouvera dans Welsh (1988), pp.1-3 un résumé des propriétés en question.

17. Logiquement, on *manque* d'autant plus d'information sur une variable (un texte) qu'une observation typique (un symbole pris au hasard) en *fournit* beaucoup.

18. Voir note 8, p. 51.

et s'interprète comme l'incertitude moyenne sur les symboles pouvant succéder à w. Si l'on note $w' \in A^{k-1}$ un k-1-gramme, l'espérance de $h(w')$ (définie pour $k \geq 1$) est appelée entropie conditionnelle *d'ordre* k et vaut:

$$h_k := E_{w'}(h(w')) = E_{w',a}(i(w' \rightarrow a)) = -\sum_{w',a} p(w'a) \log \frac{p(w'a)}{p(w')} \qquad \text{(A.16)}$$

pour $k \geq 1$ [19]. Elle correspond à l'incertitude associée en moyenne à la distribution des symboles pouvant suivre un k-1-gramme pris au hasard dans le texte, ou encore à la différence (toujours non-négative et inférieure ou égale à $\log m$ bits) entre l'entropie d'ordre k et l'entropie d'ordre k-1, en vertu de:

$$h_k = E_{w',a}(i(w' \rightarrow a)) = E_{w',a}(i(w'a)) - E_{w'}(i(w')) = H_k - H_{k-1} \qquad \text{(A.17)}$$

pour tout $k \geq 1$. La limite:

$$h_\infty := \lim_{k \rightarrow \infty} h_k \qquad \text{(A.18)}$$

est appelée *taux d'entropie* du processus et correspond à la quantité d'incertitude subsistant sur la distribution des symboles lorsque l'on dispose de toute l'information contextuelle possible, donc à l'information fournie en moyenne par un symbole selon ce modèle idéal. La non-nullité de h_∞ indique que le processus est au moins partiellement imprédictible [20].

De même que (A.15) généralise la notion d'information conditionnelle d'un symbole a étant donné un k-gramme $w \in A^k$ à l'ensemble de la distribution conditionnelle des successeurs de w, on peut définir l'information mutuelle *moyenne* de w comme l'espérance de l'information mutuelle ponctuelle $I(w \rightarrow a)$ entre w et ses successeurs:

$$I(w) := E_a(I(w \rightarrow a)) = \sum_a p(w \rightarrow a) \log \frac{p(w \rightarrow a)}{p(a)} \qquad \text{(A.19)}$$

On a toujours $I(w) \geq 0$, avec égalité ssi $(\forall\, a \in A) p(w \rightarrow a) = p(a)$. Le parallélisme s'étend à l'information mutuelle *d'ordre k*, définie pour $k \geq 1$ comme l'espérance de $I(w')$ sur tous les k-1-grammes $w' \in A^{k-1}$:

19. Pour $k = 1$, on a $h_1 = E_a\,(i(\varepsilon \rightarrow a)) = E_a\,(i(a)) = H_1$: l'entropie sur la distribution des symboles pouvant suivre la séquence vide ε est égale à l'entropie sur la distribution des symboles hors contexte.

20. Notons que le taux d'entropie n'est généralement pas connu, comme on le verra plus loin.

$$I_k := E_{w'}(I(w')) = E_{w',a}(I(w' \to a)) = \sum_{w',a} p(w'a) \log \frac{p(w'a)}{p(w')p(a)} \qquad \text{(A.20)}$$

Cette quantité s'interprète comme le gain d'information sur la distribution des symboles selon qu'on connaît (ou non) le k-1-gramme précédent, puisqu'on a:

$$I_k = E_{w,a}(I(w \to a)) = E_{w',a}(i(w' \to a)) - E_a(i(a)) = h_k - H_1 \qquad \text{(A.21)}$$

Elle est nulle ssi $(\forall\ w' \in A^{k-1}, a \in A)\ p(w'a) = p(w')p(a)$, c'est-à-dire si les k-1-grammes et les symboles sont indépendants (par construction, c'est toujours le cas de I_1).

Pour conclure ce tour d'horizon de l'application des quantités dérivées de l'entropie à la description probabiliste du texte, nous mentionnerons encore la *divergence de Kullback-Leibler* ou entropie *relative* entre deux distributions $p := p(a_1), \ldots, p(a_m)$ et $p' := p'(a_1), \ldots, p'(a_m)$ sur le même alphabet $A := \{a_1, \ldots, a_m\}$, donnée par:

$$K(p \parallel p') = \sum_a p(a) \log \frac{p(a)}{p'(a)} \qquad \text{(A.22)}$$

où l'on définit $\log \frac{0}{p'(a)} = 0$ et $\log \frac{p(a)}{0} = \infty$. L'entropie relative entre p et p' s'interprète comme la réduction d'information selon qu'on adopte le modèle p' au lieu du *vrai* modèle p. Elle est nulle ssi p et p' coïncident exactement: $(\forall\ a \in A) p(a) = p'(a)$, d'où son interprétation comme *divergence* [21] entre deux distributions.

L'entropie relative est en outre fréquemment utilisée pour l'interprétation d'autres quantités. Par exemple, l'information mutuelle moyenne étant donné $w \in A^k$ (A.19) correspond à la divergence entre la distribution des successeurs de w et celle des symboles (sans condition), donc à l'information gagnée en tenant compte du contexte w: $I(w) = K(p^w \parallel p^1)$. Dans le même ordre d'idée, l'information mutuelle d'ordre k (A.20) peut être conçue comme la divergence entre la distribution de k-grammes p^k et celle $\tilde{p}^k$ attendue sous l'hypothèse que les symboles sont indépendants du contexte $w' \in A^{k-1}$, donc une mesure de la dépendance moyenne entre un k-1-gramme et son successeur tirés au hasard: $I_k = K(p^{k-1} \parallel \tilde{p}^{k-1})$.

21. et non *distance*, puisqu'on a en général $K(p \parallel p') \neq K(p' \parallel p)$ (voir par exemple Manning & Schütze 1999, p.72, Bavaud 2000, p.264)

A.3 Aspects empiriques

Jusqu'ici, nous avons concentré notre attention sur la forme et les propriétés des modèles textuels probabilistes utilisés dans le cadre de l'ASTC; nous n'avons évoqué que des modèles abstraits ou dont les paramètres ont été spécifiés arbitrairement en fonction des besoins didactiques de l'exposé. En pratique, on est plus fréquemment conduit à *estimer* les paramètres, c'est-à-dire à leur attribuer des valeurs observées dans un échantillon, ou à les *inférer*, c'est-à-dire à se servir des données pour sélectionner un modèle parmi ceux résultant de la variation des paramètres. Dans le cas d'une chaîne de Markov, il s'agit non seulement d'estimer les probabilités de transitions $\hat{p}(w \rightarrow a)$ [22] pour tout $w \in A^k$ et $a \in A$ mais aussi - et même surtout - d'inférer l'ordre $\hat{k}$ du processus.

L'estimation des probabilités de transition implique préalablement celle des probabilités $p(w)$ constituant les distributions de k-grammes p^1,p^2,... Etant donné un corpus (au sens très général de séquence observée) $\sigma := s_1^n := s_1 \dots s_n$, la probabilité d'un k-gramme $w \in A^k$ est estimée par la *fréquence relative* correspondante $f(w)$, soit le rapport entre le *nombre d'occurrences* $n(w)$ (ou *effectif*, ou encore fréquence *absolue*) du k-gramme et le nombre total de $k-$grammes:

$$\hat{p}(w) := f(w) := \frac{n(w)}{n-k+1} \tag{A.23}$$

L'estimation des probabilités de transition d'ordre k est donnée par:

$$\hat{p}(w \rightarrow a) := \frac{n(wa)}{\sum\limits_{\tilde{a} \in A} n(w\tilde{a})} \tag{A.24}$$

et dépend, comme on le voit, de la distribution des k+1-grammes [23].

L'ordre $\hat{k}$ du processus [24] peut être inféré sur la base des propriétés informationnelles du corpus. Ainsi qu'on l'a vu plus haut, l'entropie conditionnelle

22. Ici et dans ce qui suit, l'usage du circonflexe explicite le caractère *estimé* ou *empirique* des quantités considérées (*inféré*, s'il s'agit de l'ordre $\hat{k}$).

23. L'expression (A.24) est le pendant empirique de (A.3); leur différence traduit l'une des «anomalies» du formalisme, à savoir que l'on n'a pas toujours $n(w) = \Sigma\, n(w\tilde{a})$, puisque le dernier k-gramme d'un texte n'a pas de successeur. Toutefois, pour un échantillon suffisamment grand, on peut raisonnablement considérer que $n(w)$ est une bonne approximation de $\Sigma\, n(w\tilde{a})$, de même que n pour $n-k$+1 dans (A.23).

24. Il s'agit ici de l'ordre d'une chaîne de Markov, donc la taille k du contexte w qui conditionne les probabilités de transition; dans ce qui suit, il importe de distinguer l'ordre ainsi défini de celui de l'entropie conditionnelle h_k, qui correspond à celui de la distribution de k-grammes à partir de laquelle une chaîne de Markov d'ordre k-1 peut être spécifiée.

d'ordre k $\hat{h}_k$ est une mesure de l'incertitude moyenne sur la distribution des symboles étant donné le k-1-gramme précédent. Cette quantité est d'autant plus faible que le modèle fournit beaucoup d'information sur le processus, et l'on s'attend à ce qu'elle décroisse en fonction de k jusqu'à l'ordre $\tilde{k}$ où elle atteint son minimum, appelé taux d'entropie du processus h_∞ (A.18), si le processus est d'ordre $\tilde{k}-1$. En ce sens, le taux d'entropie constitue une forme d'idéal informationnel dont il est naturel de vouloir s'approcher.

Mais $\tilde{k}$ peut être arbitrairement élevé, et à moins que le taux d'entropie soit nul, on ne peut en pratique jamais s'assurer que la décroissance de $\hat{h}_k$ est bel et bien terminée[25]. Même si h_∞ est nul, une chaîne de Markov d'ordre $\tilde{k}-1$ peut être suffisamment complexe (en terme du nombre de ses paramètres) pour qu'on lui préfère un modèle d'ordre inférieur. On adopte alors une stratégie de sélection basée sur l'observation de la différence $\hat{d}_k := \hat{h}_k - \hat{h}_{k-1}$ ou entropie *résiduelle* d'ordre k définie pour $k \geq 1$, qui s'interprète comme la réduction d'incertitude selon qu'on connaît le k-gramme précédent plutôt que le k-1-gramme seulement (Bavaud 1998, p.213); on a toujours $\hat{d}_k \geq 0$, avec égalité ssi l'entropie conditionnelle d'ordre k est égale à celle d'ordre $k+1$, c'est-à-dire si la croissance de l'entropie sur les distribution de k-grammes $\hat{p}^{k-1}$,$\hat{p}^k$ et $\hat{p}^{k+1}$ est linéaire:

$$\hat{d}_k = 0 \Leftrightarrow \hat{h}_k = \hat{h}_{k+1} \Leftrightarrow \hat{H}_k = \frac{\hat{H}_{k-1} + \hat{H}_{k+1}}{2} \qquad \text{(A.25)}$$

Autrement dit, la nullité de l'entropie résiduelle à l'ordre k indique qu'on ne gagne aucune information en estimant une chaîne de Markov à partir de $\hat{p}^{k+1}$ plutôt que de $\hat{p}^k$. Inversement, une valeur élevée de $\hat{d}^k$ signifie qu'une chaîne d'ordre k fournit significativement plus d'information sur le processus qu'une chaîne d'ordre k-1.

Pour illustrer ce raisonnement, considérons l'exemple d'un texte codé sur $A := \{0,1\}$. Supposons qu'il ait été généré par un processus parfaitement aléatoire, avec $p(0) = p(1) = 0.5$. Sous l'hypothèse d'indépendance des états successifs, on s'attend à ce que les distributions de k-grammes soient uniformes pour $k = 1,2,\ldots$, et l'entropie associée $\hat{H}^k$ maximale et égale à $\log m^k = k$ bits : $\hat{H}_1 = 1$ bit, $\hat{H}_2 = 2$ bits, ... d'où une entropie conditionnelle $\hat{h}^k$ constante et maximale: $\hat{h}^1 = \hat{h}^2 = \ldots = 1$ bit (toutes les distributions conditionnelles sont uniformes) et une entropie résiduelle constante et minimale: $\hat{d}^1 = \hat{d}^2 = \ldots = 0$ bit (aucune réduction d'incertitude n'est possible).

Par contraste, admettons que le texte ait été généré par un processus d'ordre 2, avec $p(00 \to 1) = 1$, $p(01 \to 1) = 1$, $p(11 \to 0) = 1$, $p(10 \to 0) = 1$ et

25. pour des raisons de représentativité que nous évoquerons plus loin

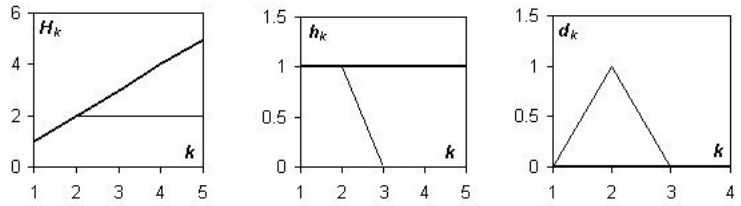

FIG. A.1 – *entropogrammes attendus pour un processus binaire déterministe d'ordre 2; la courbe en gras représente le cas d'un processus aléatoire.*

toutes les autres probabilités de transition nulles: ...0011001100110011... Les symboles et les bigrammes sont donc équirépartis, d'où $\hat{H}_1 = 1$ bit et $\hat{H}_2 = 2$ bits; les distributions d'ordre supérieur à 2 sont également uniformes, mais sur quatre k-grammes seulement au lieu des 2^k possibles: $\hat{H}_3 = \hat{H}_4 = \ldots = \hat{H}_2 = 2$ bit. L'entropie conditionnelle est donc maximale pour $k = 1$ et 2, puis nulle étant donné que la probabilité des symboles n'est pas plus affectée par un contexte de taille $k \geq 3$ qu'elle ne l'est par le seul bigramme précédent. Enfin, l'entropie résiduelle est toujours nulle sauf à l'ordre 2, où l'on observe un maximum: $\hat{d}_1 = 0$, $\hat{d}_2 = 1$, $\hat{d}_3 = \hat{d}_4 = \ldots = 0$; seul le passage de la distribution des bigrammes à celle des trigrammes (donc d'une chaîne d'ordre 1 à une chaîne d'ordre 2) permet de réduire l'incertitude du modèle - en l'occurrence de l'annuler complètement [26].

Cette procédure inférentielle peut être visualisée graphiquement à l'aide des *entropogrammes* de Bavaud ou graphes des entropies (de Shannon, conditionnelle et résiduelle d'ordre k) rapportées à l'ordre (Bavaud 1998, p.213). Les entropogrammes attendus pour notre exemple sont représentés sur la figure A.1 (page suivante), où l'on voit clairement que $\hat{H}_k$ est linéaire *jusqu'à* et *à partir de* l'ordre 2, mais pas sur l'intervalle $k = 1, 2, 3$ comme on a $H_2 > \frac{H_1+H_3}{2}$. Il résulte de cette variation de la pente de $\hat{H}_k$ que $\hat{h}_k$, constante jusqu'à l'ordre 2 et à partir de l'ordre 3, décroît de 1 bit entre les deux; on retrouve cette quantité sur le graphe de $\hat{d}_k$, sous la forme d'un pic à l'ordre 2.

La régularité des résultats témoigne du caractère déterministe du processus modélisé. En fait, dans ce cas extrême, on parviendrait à des résultats comparables en utilisant au lieu de l'entropie la *variété* d'ordre k, définie (ici empiriquement) comme le nombre de k-grammes *différents* observés dans le texte:

26. Dans ce cas, $\tilde{k} = 2$.

$\hat{v}_k :=| \{w \in A^k | n(w) \geq 1\} |$, et en dérivant les variétés conditionnelle et résiduelle correspondantes. Par contre, un test basé sur la variété échouerait à détecter l'ordre d'un processus d'ordre $k \geq 1$ plus proche de l'indépendance. Dans ce cas, $\hat{d}_k$ serait une valeur quelconque comprise entre 0 et 1, dont il s'agit de déterminer si elle est *significativement* éloignée de 0, indiquant qu'un modèle d'ordre k est *significativement* meilleur qu'un d'ordre k-1 du point de vue de l'information qu'il fournit sur le processus (voir figure A.2 p. 61).

La significativité de $\hat{d}_k$ est d'autant plus élevée que le corpus ayant servi à estimer les entropies est grand et le nombre de paramètres à l'ordre k petit. Cela implique que, pour qu'on considère que son éloignement de zéro n'est pas lié à la mauvaise qualité de son estimation, elle doit être d'autant plus élevée que k est grand. Les deux contraintes conditionnent la forme définitive du *test de l'ordre du processus* [27] opposant:
$H_0(k)$: «Le processus est d'ordre k»;
$H_1(k)$: «Le processus est d'ordre k+1»
On rejette $H_0(k)$ au niveau α ssi

$$2 \ln 2 (n-k) d_{k+1} \geq \chi^2_{1-\alpha}[m^k (m-1)^2] \tag{A.26}$$

où $n-k$ représente le nombre de k+1-grammes observés dans le texte et $m^k(m-1)^2$ le nombre de paramètres *libres* [28] d'une chaîne de Markov d'ordre k pour l'alphabet considéré [29]. On commence par $k = 0$ et le test est réitéré pour k = 1,2,... On sélectionne finalement l'ordre k le plus élevé induisant une réduction significative d'incertitude et pour lequel on dispose d'une estimation *fiable* des probabilités de transition d'ordre k.

La fiabilité de l'estimation $\hat{p}(w \to a)$ définie par (A.24) est directement liée à la notion de *représentativité* du corpus, dont nous avons déjà différé la discussion par deux fois. Précisons à ce sujet que la longueur n du corpus et la taille m de l'alphabet imposent une limite absolue à la portée de nos estimateurs. Nous avons déjà mentionné plus haut [30] l'existence de l'*effet de bord*, c'est-à-dire que le nombre de k-grammes observés dans un texte de longueur n décroît d'une unité à chaque ordre; à l'extrême, ce texte ne contient qu'un seul n-gramme, vraisemblablement insuffisant pour estimer les $m^{n-1}(m-1)^2$ paramètres d'une

27. Voir par exemple Bavaud 1998, pp.214-5.
28. Une distribution de m probabilités a m-1 paramètres libres, la dernière étant contrainte de porter le total à 1.
29. Il est à noter que la présence du facteur ln 2 est liée à l'utilisation du logarithme binaire, et qu'il disparaîtrait si l'on travaillait en base e; en outre, le facteur $n-k$, peut être remplacé par n pour n grand (voir note 23, p. 57).
30. Voir note 23, p.57.

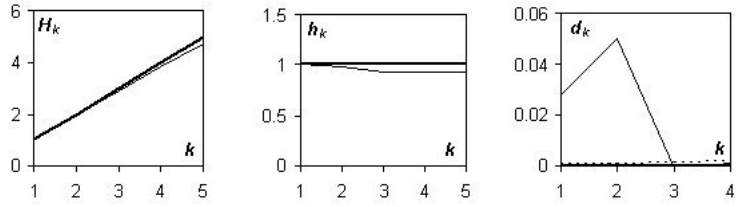

FIG. A.2 – *Entropogrammes observés sur une séquence binaire de 50'000 symboles générée par un processus non-déterministe d'ordre 2*[31]*; la courbe en traitillé représente le seuil de significativité de* $\hat{d}_k$ *au niveau* $\alpha = 0.001$

chaîne de Markov d'ordre n-1. On peut montrer que l'estimation des probabilités de transitions d'ordre k est *fiable* tant que l'est celle de la distribution de k+1-grammes correspondante, c'est-à-dire (au moins [32]) tant que:

$$k + 1 \leq \log_m n = \frac{\log n}{\log m} \tag{A.27}$$

Le test (A.26) connaît la même limite de fiabilité, et il se peut que l'effet de bord biaise le comportement des indicateurs avant qu'on puisse rejeter $H_0(k)$, auquel cas on sera contraint de sélectionner le modèle d'ordre 0 (A.4).

Il est à noter que la représentativité du corpus n'est pas évaluée ici en fonction du processus de sélection de l'échantillon, comme c'est le cas en psychologie notamment, où un échantillon est dit *représentatif d'une population* ssi chaque individu de cette population a la même chance d'y apparaître. Il est particulièrement délicat de justifier la nature de la population dont serait issu un texte (au sens traditionnel, comme l'œuvre intégrale d'un auteur, par exemple) et, partant, celle du processus ayant abouti à sa sélection. On préfère généralement considérer que l'échantillonnage est *exhaustif*, et que les données observées correspondent à l'ensemble des données observables jusqu'alors. Revenant à nos considérations initiales de l'annexe A.1, nous pouvons dire plus précisément que la justification d'une modélisation probabiliste se trouve dans le caractère *in-*

31. spécifié par les probabilités $p(00 \to 1) = 0.75$, $p(01 \to 1) = p(10 \to 1) = 0.5$ et $p(11 \to 1) = 0.25$ (les autres probabilités peuvent être déduites aisément)

32. Pour être précis, l'estimation est bonne tant que $k+1 \leq \log n/h_\infty$ (Shannon 1948), mais h_∞ n'est généralement pas connu; en revanche , on sait que $h_\infty \leq \log m$, d'où $\log n/\log m \leq \log n/h_\infty$ (Bavaud 1998, p.214).

trinsèquement aléatoire que l'on prête au processus textuel décrit, et que le formalisme que nous achevons d'introduire traduit par la non-nullité du taux d'entropie du processus[33].

33. Un lecteur parvenu jusqu'à ce point appréciera sans doute l'existence du freeware Entropizer 1.1 (pour Macintosh PowerPC, téléchargeable librement sur le site de la section de linguistique de l'UNIL (http://www.unil.ch/ling).

Annexe B

Evaluations globales [1]

Abréviations: v = variété conditionnelle, h = entropie conditionnelle, I = information mutuelle moyenne, K = divergence entre distribution conditionnelle et initiale, oX = ordre X.

1. Nous rappelons que tous les résultats indiqués dans cette partie sont produits par les indicateurs *moyens* (voir alinéa 2.2.2.3, équation (2.11)), ce que nous ne répéterons plus dans la suite.

B.1 Taux d'erreur

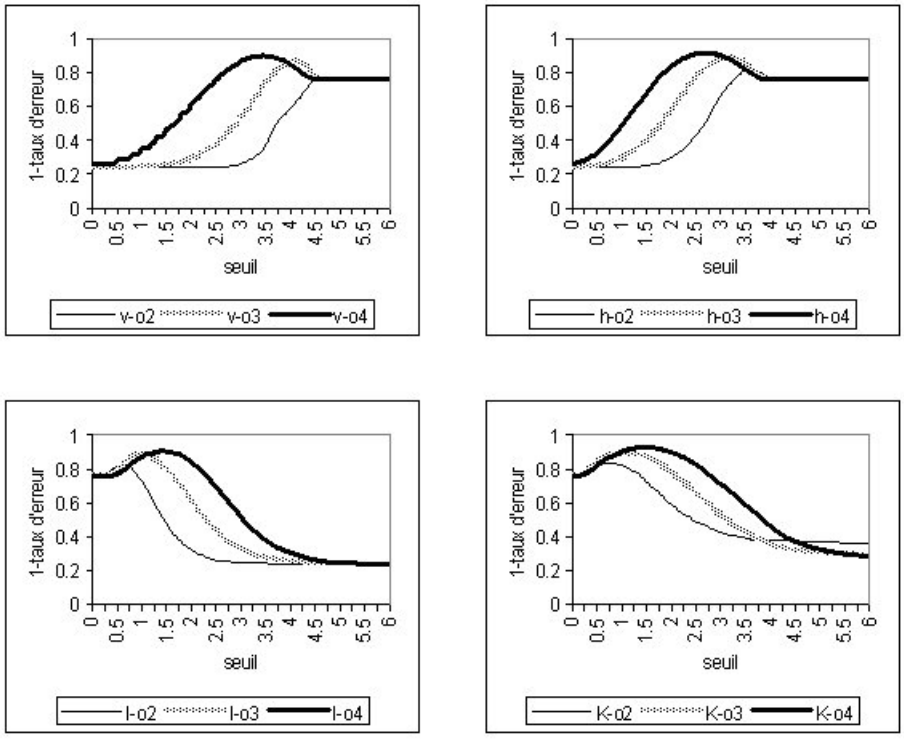

FIG. B.1 – *Taux d'erreur pour chaque critère et à chaque ordre (corpus AUSTEN)*

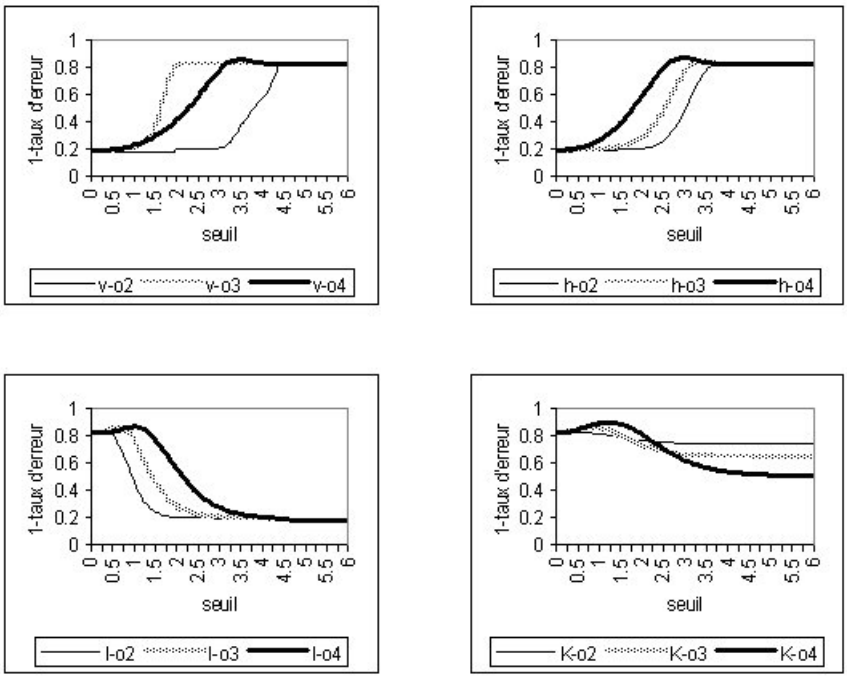

FIG. B.2 – *Taux d'erreur (corpus OVIDE)*

B.2 Courbes ROC

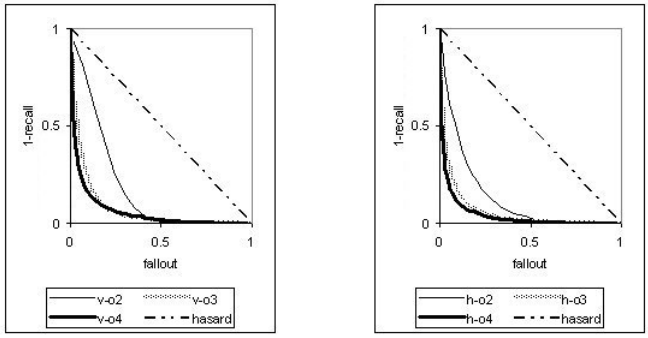

FIG. B.3 – *courbe ROC pour chaque critère et à chaque ordre (corpus AUSTEN)*

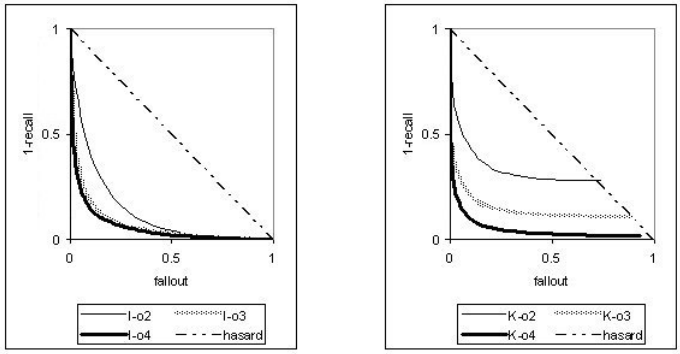

FIG. B.4 – *courbes ROC (corpus AUSTEN) - suite*

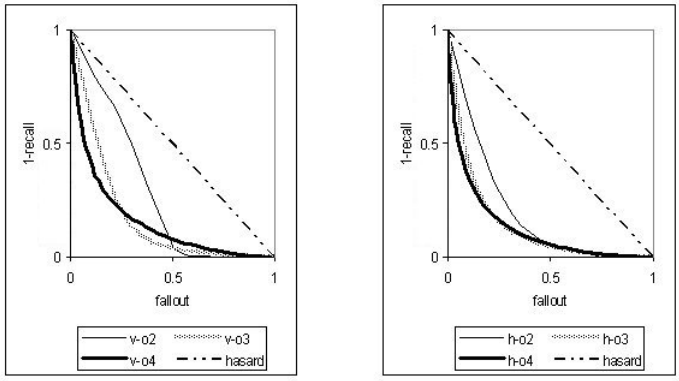

FIG. B.5 – *courbes ROC (corpus OVIDE)*

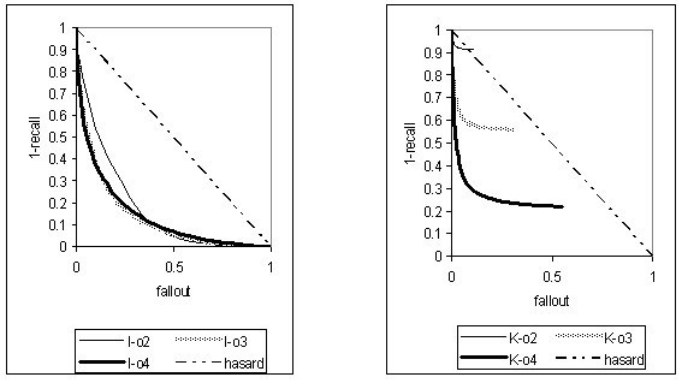

FIG. B.6 – *courbes ROC (corpus OVIDE) - suite*

Annexe C

Extraits de corpus segmentés [1]

Rappel: le symbole '=' indique une frontière manquée, et '_' une fausse alarme.

Abréviations: v = variété conditionnelle, h = entropie conditionnelle, I = information mutuelle moyenne, K = divergence entre distribution conditionnelle et initiale, oX = ordre X, sY = seuil Y, f = fallout (fausse alarmes), 1-r = 1-recall (manqués), e = taux d'erreur [1].

C.1 Corpus AUSTEN

La proportion d'espaces dans ce corpus est de 19.2%. En n'insérant aucune frontière, on obtient un taux d'erreur de 23.7%.

o4-v-s3.45: woodhouse hand_some clever and rich with a comfor_t_able home and happy disposition seem_ed to unite some of the best b_les_s_ing_s of=existence and had lived nearly twenty=one year_s in the=world with very little to distress=or=vex=her=s_he w_a_s the young_est of the two daught_er_s of a most affection_at_e indulgent=fathe_r and had in consequence of her s_ister s marriage been mistress of his house from a very early period=her mo_ther had=di_ed=to_o long=ago=for her to have mor_e than an in_distinct=remembrance of her caress_e_s and her place had been supplied by an=excellent woman=as=gover_ness who=had fallen little short of=a mo_ther in affection
e: 10.1%, 1-r: 23.6%, f: 5.9%

1. Voir note 32, p. 61.

1. Les trois dernières valeurs sont exprimées ici en pourcentages.

o4-h-s2.72: woodhouse h_and_some clever and rich with a comfort_able ho_me and happy disposition seem_ed to unite some of the best b_les_s_ing_s of existence and had lived nearly twenty=one year_s in the=world with very little to distress=or=vex=her=s_he was the you_ngest of the two daught_er_s of a most affection_at_e indulgent=fa_ther and had in consequence of her sister s marriage been mistress of his house from a=very early period=her mo_ther had di_ed=to_o long=ago=for=her to have more than an in_distinct=remembrance of her caress_e_s and her place had been supplied by an=excellent woman=as=gover_ness who had=fallen=little short of a mo_ther in affection
e: 8.4%, 1-r: 20.6%, f: 4.6%

o4-I-s1.36: woodhouse h_and_some clever=and rich with a comfort_able home and happy disposition seem_ed=to unite=some of the best b_les_s_ing_s of=existence and had lived nearly twenty=one year_s in=the=world with very little to distress=or=vex=her=s_he was the youngest of the two daughter_s of a most affection_at_e indulgent=fa_ther and had in consequence of her sister s marriage been mistress of h_is=house from a=very early=period=her m_o_ther had=died=too long=ago=for=her to have more than an=in_distinct=remembrance of her caress_es and her place had been supplied=by an=excellent woman=as=gover_ness who had=fallen=little short of=a=mo_ther in affection
e: 9.4%, 1-r: 26.9%, f: 4%

o4-K-1.38: woodhouse h_and_some clever and rich with a comfort_able home and happy disposition seemed to unite some of the best=bles_s_ing_s of existence and had lived nearly twenty=one year_s in the world with very little t_o distress=or=vex=her=s_he was the youngest of the two daughter_s of a most affection_ate indulgent=father and had in consequence of her sister s marriage been mistress of his house from a=very early=period=her mother had died too long=ago for her to have more than an=in_distinct=remembrance of her caresses and her place had been supplied by an=excellen_t woman=as=gover_nes_s who had=fallen=little short of a=mother in affection
e: 7.4%, 1-r: 22.9%, f: 2.6%

o3-v-s4.14: a=woodhous_e hand_some cle_ver and rich with a comfor_t_able home and happy=disposition s_e_em_ed to=unit_e some of the best=ble_s_s_ing_s=of=exist_ence and had lived=near_ly twenty=one=year_s in t_he world=with very little to dist_res_s=or=vex=her s_he was the young_e_s_t of the two=daugh_t_er_s of a=most=affection_at_e indulgent=fathe_r and had in cons_equence of her s_is_ter s marriage=been mist_res_s of his hous_e from

a=very early period=her mother had=died too long=ago=for her to have more than an=in_d_i_st_inct=remembrance of her caress_es and her place had be_en supplied=by=an=excell_ent= wo_man=as=gover_nes_s who had= fallen=little short o_f a=moth_e_r in=affection
e: 12.8%, 1-r: 28%, f: 8.1%

o3-h-s3.21: a=woodhouse=hand_some clever=and rich with a=comfor_table home and happy disposition se_em_ed to unit_e some of the best=bles_s_ing_s of=exist_ence and had lived near_ly twenty=one year_s in the world with=very little to di_stres_s=or=vex=her she was the younges_t of the two=daughter_s of a most affection_ate indulgent=fa_the_r and had in cons_equence of her s_ister s marriage=been mistres_s of his house from=a=very ear_ly period=her mo_ther had=di_ed to_o long ago=for=her to have more than an=ind_is_tinct=remembrance of her caress_es and her place had=been supplied by=an=excell_ent wo_man=as=governess who had=fallen=little short o_f a mother in=affection
e: 10.8%, 1-r: 26.2%, f: 6%

o3-I-s0.96: a=woodhouse=hand_some clever=and rich with=a= comfor_t_able=home and happy disposition seemed to=unite some of the best=ble_s_s_ing_s of=exist_ence and had lived near_ly twent_y=one year_s in the world with very little=to di_st_res_s=or=vex=her she was the younges_t of the two=daughter_s of a=most affection_ate=indulgent=fathe_r and had in cons_equence of her s_ister s marriage=been mist_res_s of his=hous_e from=a=very=ear_ly period=her mo_ther had=died too long ago=for her to have more than=an=ind_i_st_inct=remembrance of her care_ss_es and her place had=be_en supplied=by=an=excell_ent woman= as=governess who=had=fallen=little=short o_f a=mother in=affection
e: 11.2%, 1-r: 29.7%, f: 5.4%

o3-K-1.04: a=wood_house=hand_some=clever=and rich with=a= comfortable=home=and happy disposition seemed to unite=some=of the=best=ble_s_s_ing_s of=existence and had lived near_ly twent_y one=year_s in the=world with very little to distres_s=or=vex=her s_he was the=younges_t of the=two=daughter_s of a=most affectionate=indulgent father and had=in consequence of her s_is_ter s marriage been mistres_s of his house from=a=very=ear_ly period=her mother had=died too long ago=for her to have more=than=an= ind_istinct=remembrance of her caresses and her place had been supplied by=an=excellent woman=as=governess who

had=fall_en=little=s_hort o_f a=mother in=affection
e: 10.5%, 1-r: 33.4%, f: 3.4%

o2-v-s4.6: ma=woodhouse=handsome=clever and=rich=with=a=comfortable=home=and=happy=disposition=se_emed=to=unite=some=of=the=best=blessings=of=existence=and=had=lived=ne_arly=twenty=one-years=in=the=world=with=very=little=to=distress=or=vex=her=she=was=the=youngest=of=the=two=daughters=of=a=most=affectionate=indulgent=fatherand=had=in=consequence=of=her=sister=s=marriage=been=mistress=of=his=house=from=a=very=early=period=her=mother=had=died= too=long=ago=for=her=to=have=more=thananindist_inct= remembrance=of=her=caressesand=her=place=had=been=supplied=by=an=excellent=woman as=governess=who=had=fallen=little=short=of=a=mother=in=affection
e: 23.4%, 1-r: 95.8%, f: 9%

o2-h-s3.54: ma=woodhou_se hand_some clever and rich with=a=comfor_table hom_e and happy=dis_position s_e_em_ed to un_it_e som_e of the be_st=bles_s_ing_s=o_f exis_t_ence and had lived ne_arly twent_y=one y_ears in the world with=very little to dis_t_res_s=or=vex=her she was the young_es_t=of the=two=daughter_s=of a most affection_at_e indulgent=fa_ther and had in=con_se_quence of=her sis_ter s marriage be_en mis_t_res_s of=his hou_se fro_m=a=ver_y early period=her mother had di_ed to_o=long ago=for=her to have=mor_e than=an=ind_is_t_inct= re_mem_brance of=her car_es_s_e_s and=her place had be_en suppli_ed=by an=excell_ent=wom_an=as gover_ness who=had fall_en little shor_t=o_f a mothe_r in=affection
e: 18.1%, 1-r: 35.9%, f: 12.5%

o2-I-s0.63: ma=woodhouse hand_some=clever and rich with=a=com_for_table=home and happy=disposition se_emed to=unite=some of the be_st=bles_s_ing_s of exis_t_ence and had lived nearly twent_y=one=year_s in=the world with=very little to dist_res_s=or vex=her she was the young_e_s_t of the=two=daughter_s of=a=most affection_ate=indulgent=father and had in=con_se_quence of=her sis_ter s marriage be_en mist_res_s of=his=house from=a=ver_y=early=period=her mother had died too=long ago=for=her to have=mor_e than=an=ind_is_t_inct=re_mem_brance of=her car_es_s_e_s and=her place had be_en supplied=by an=excell_ent wom_an=as gover_nes_s who=had fall_en little shor_t=of=a=mother in=affection
e: 16.8%, 1-r: 42.3%, f: 8.9%

o2-K-0.77: woodhouse h_and_some clever and rich with a comfort_able home and happy disposition seemed to unite some of the best=bles_s_ing_s of existence and had lived nearly twenty=one year_s in the world with very little t_o distress=or=vex=her=s_he was the youngest of the two daughter_s of a most affection_ate indulgent=father and had in consequence of her sister s marriage been mistress of his house from a=very early=period=her mother had died too long=ago for her to have more than an=in_distinct=remembrance of her caresses and her place had been supplied by an=excellen_t woman=as=gover_nes_s who had=fallen=little short of a=mother in affection
e: 16.4%, 1-r: 55.8%, f: 4.1%

C.2 Corpus OVIDE

La proportion d'espaces dans ce corpus est de 14.9%. En n'insérant aucune frontière, on obtient un taux d'erreur de 17.5%.

o4-v-s3.55: va fert animus mutatas dicere formas=corpora=di coeptis= nam=vos=mutastis=et illas= adspirate=me_is=prima_que=ab= originem_undi= ad=mea= perpetuum=de_ducite t_empora carmen ante mare et terras et quod tegit omni_a caelum=unus erat toto natur_a_e vult_us=in orbe=quemdixere=chaos=rudis=indigesta_que= moles=nec quicquam nisi pondus=iners=congesta_que=eodem=non=bene=iunctarum discordia= semina=r_erum nullus=adhuc=mundo= praebebat lumin_a t_itan=nec=nova crescendo reparabat cornua=phoebe=nec circumfuso=pendebat in a_ere tellus ponderibus=librata=suis=nec bracchia longo=margine terrarum=porrexerat
e: 14.6%, 1-r: 59.3%, f: 5.1%

o4-h-s3.06: va fert=animus mutatas dicere formas= corpora=di coeptis=nam=vos=mutastis et illa_s= adspirate=meis=prima_que=ab=origine m_undi=ad=me_a= perpetuum=de_ducite tempora carmen=ante mare et terras et quod tegit omnia caelum=unus erat toto=natur_a_e vultus=in orbe=quem dixere=chaos=rudis indigesta_que moles=nec quicquam nisi pondus=iners=congesta_que=eodem=non bene=iunctarum discordia=semina r_erum nullus=adhuc mundo=praebebat lumina t_itan=nec=nova crescendo reparabat cornua phoebe nec circumfuso=pendebat in a_ere=tellus ponderibus=librata=su_is=nec bracchia longo=margine terrarum porrexerat
e: 13.3%, 1-r: 47%, f: 4.1%

o4-I-0.99: va fert animus mutatas dicere formas=corpora=di coeptis=nam=vos=mutas_tis et illa_s adspirate=meis= prima_que=ab=origine m_undi=ad=me_a=per_petuum= de_ducit_e tempora=carmen=ante mare et terra_s et quod tegit omnia=caelum=unus erat toto=n_a_tura_e vultus= in=orbe=quem dixere=chaos=rudis=indigesta_que moles=nec quicquam nisi pondus=iners=congesta_que=eodem=non= bene=iuncta_rum discordia=semina=rerum nullus=adhuc mundo praebebat lumina t_itan nec=nova crescendo reparabat cornua=phoebe nec circumfuso=pendebat in=a_ere=tellus ponderibus=librata=suis=nec bracchia longo=margine terrarum=porrexerat
e: 13.1%, 1-r: 57.3%, f: 3.8%

o4-K-1.14: va=fert animus mutatas=dicere formas corpora=di=coeptis= nam=vos=mutastis=et=illas adspirate=meis= prima_que=ab=origine mundi=ad=me_a=per_petuum de_ducite=tempora carmen=ante mare et terras=et quod tegit omnia caelum=unus=erat toto=na_turae vultus in orbe quem=dixere=chaos=rudis=indigesta_que moles nec quicquam nisi pondus=iners=con_gesta_que eodem non bene iuncta_rum discordia semina rerum nullus adhuc mundo praebebat=lumina titan nec nova crescendo re_parabat cornua=phoebe nec circumfuso=pendebat=in=aere tellus ponderibus librata suis=nec bracchia longo margine terrarum porrexerat
e: 10.5%, 1-r: 46%, f: 3%

o3-v-s3.81: ova fer_t a_n_im_us muta_t_as dice_re formas cor_pora d_i coeptis nam=vos=muta_s_ti_s et illa_s ad_spira_t_e=m_e_is=prima_que ab=origine m_undi ad=mea per_petu_um=de_ducit_e t_em_pora carmen ant_e m_are et ter_ras et quod tegit omni_a ca_e_lum unus e_r_at tot_o=na_t_u_r_a_e vult_us in orbe=quem dixere=chaos= rudi_s in_di_gest_a_que m_oles nec quic_quam ni_s_i pondus i_n_ers=congest_a_que=e_odem non bene iunc_ta_r_um dis_cordi_a=s_e_m_i_n_a=r_e_rum nullus=adhuc mundo praebebat lum_i_n_a t_i_tan nec nova cres_cendo r_e_p_arabat cornua=phoebe nec circum_fuso= pendebat i_n a_e_r_e t_ellus ponde_ribus libra_t_a=s_u_is nec bracchia longo=margine t_er_rar_um porrex_era_t
e: 23%, 1-r: 23.3%, f: 23%

o3-h-s3.4: ova=fert=animus mutat_as dicere formas corpora di coeptis nam vos mutas_tis et illas=ad_spirat_e=me_is= prima_que=ab=origine m_undi=ad=me_a=perpetuum= de_ducit_e=tempora carmen ante=m_are et terras et quod tegit=omnia caelum=unus erat toto=natur_ae vultus in=orbe=quem=dixere=chaos=rudis in_digest_a_que moles nec quicquam

nis_i=pondus=iners=congest_a_que=eodem non=bene iunctarum
discordi_a=s_emina=r_erum nullus= adhuc mundo praebebat lumina
t_i_tan=nec=nova= crescendo=reparabat cornua=phoebe=nec circum_fuso=
pendebat=in a_e_re tellus=ponderibus=librat_a=su_is= nec=bracchia
longo=margine terrarum=porrexerat
e: 15.5%, 1-r: 59.1%, f: 6.3%

o3-I-0.62: ova=fer_t=animus mutatas dicere formas corpora di=coeptis
nam=vos mutas_tis et illas adspirat_e=meis= primaque=ab=origine
mundi=ad=mea=per_petuum= deducit_e=tempora carmen=ante=mare=et
terras=et quod tegit=omnia=caelum=unus erat toto=natur_ae=vultus=
in=orbe=quem dixere=chaos=rudis=in_digest_aque moles nec
quicquam=nisi= pondus=iners=congest_aque=eodem non=bene=
iunctarum=discordia= semina=rerum nullus=adhuc mundo praebebat=
lumina ti_tan=nec= nova cres_cendo=reparabat cornua=phoebe=nec
circumfuso= pendebat=in=aere tellus ponderibus=librat_a=suis
nec=bracchia=longo=margine terrarum=porrexerat
e: 13.8%, 1-r: 63.3%, f: 3.3%

o3-K-0.78: ova=fert=animus mutatas=dicere formas= corpora=di=coeptis
nam vos mutastis=et=illas adspirate= meis=prima_que=ab=origine
mundi=ad=mea=per_petuum de_ducite=tempora carmen=ante
mare=et=terras=et=quod tegit omnia=caelum=unus=erat=toto=naturae
vultus=in= orbe=quem=dixere=chaos=rudis=in_digesta_que moles nec
quic_quam=nisi=pondus=iners=congesta_que eodem=non=bene=iunctarum
discordia=semina=rerum nullus adhuc mundo praebebat=lumina=titan=
nec=nova crescendo= reparabat=cornua=phoebe=nec circumfuso=
pendebat=in=aere tellus=ponderibus=librata=suis nec bracchia=longo
margine=terrarum=porrexerat
e: 14.2%, 1-r: 70.1%, f: 2.4%

o2-v-s4.15: _nova=fer_t a_n_im_us=mut_a_t_as dic_e_r_e= for_ma_s cor_por_a
d_i c_o_eptis=nam=vos=mut_a_s_t_i_s e_t il_la_s adspi_r_a_t_e m_e_is
pri_m_aque ab=o_r_igin_e=m_undi ad me_a per_pe_tu_um=deduc_i_t_e
t_em_por_a c_ar_men an_t_e m_a_re et ter_ra_s et quod=tegit omni_a=
c_a_e_l_um unu_s e_r_at to_t_o=n_a_t_u_r_a_e=vultu_s i_n orbe=quem
dix_er_e=c_ha_os ru_di_s in_di_ges_t_a_que m_o_les nec=quicquam ni_s_i
pondus i_n_er_s=con_ges_t_a_que e_od_em non=be_n_e i_uncta_r_um
di_s_cor_di_a=s_e_m_in_a r_e_r_um nullus adhuc=mundo pra_ebebat lum_in_a
t_it_an nec nova=c_re_s_cendo r_e_pa_r_abat cor_nu_a=phoebe

n_ec=circumfus_o pendebat in a_e_r_e t_ellus ponde_r_ibus librat_a s_u_is
nec=bracchi_a l_ongo=margin_e t_er_ra_r_um por_rex_er_at
e: 36.9%, 1-r: 30%, f: 38.5%

o2-h-s3.41: _nova fert=a_nimus=mut_a_t_as di_c_ere formas cor_por_a d_i
c_oeptis nam vos=mut_as_t_i_s e_t illa_s adspira_t_e=me_is pri_ma_que
ab=or_igin_e=m_undi ad=mea p_erpetuum de_duc_i_t_e=t_em_p_or_a c_armen
ant_e=m_are et terra_s et quod=tegit omni_a c_a_e_lum unus e_r_at
toto=n_a_tur_a_e vultus i_n orbe=quem dixer_e=chaos rudi_s in_di_ges_ta_que
m_oles=nec=quicquam ni_s_i=pon_dus in_ers=con_ges_ta_que eodem
non=bene i_unc_ta_r_um dis_cordi_a s_e_m_in_a=r_e_r_um nul_lus adhuc=
mundo pra_ebebat lum_in_a t_i_t_an=nec nova=crescendo= r_e_parabat
cor_nua=phoebe=nec=circum_fus_o=p_endebat in a_e_r_e t_ellus ponde_ribus
libra_t_a su_is nec=bracchi_a longo margin_e t_erra_r_um p_orrex_er_a_t
e: 26.8%, 1-r: 26.2%, f: 26.9%

o2-I-0.39: _nova=fer_t=animus=mutatas dicere= for_mas
cor_pora=di=coeptis=nam=vos=mutas_tis et=illas adspirate=meis
primaque=ab=origine=mundi=ad=mea=per_petuum= deducit_e=tempora
=car_men =ante=mare=et terras et quod=tegit=omnia=caelum=unus=
erat=toto=naturae=vultus=in=orbe=quem=dixere=chaos=rudis=
in_digestaque moles=nec=quicquam=nisi=pondus=iners
=con_gestaque=eodem=non=bene=iunc_tarum=dis_cor_dia=semina=
rerum=nullus=adhuc=mundo=praebebat =lumina=titan=nec
nova=crescendo=reparabat cor_nua=phoebe=nec= circumfuso=pendebat=
in=aer_e=tellus ponderibus=librata suis=nec=bracchia=longo=
margin_e=terrarum=porrexerat
e: 16.1%, 1-r: 80.5%, f: 2.5%

o2-K-0.81: nova=fert=animus mutatas=dicere=formas=
corpora=di=coeptis=nam=vos=mutastis=et=illas= adspirate=meis=
prima_que=ab=origine=mundi=ad=mea=perpetuum=de_ducite=
tempora=carmen=ante=mare=et=terras=et=quod=tegit=omnia=caelum=
unus=erat=toto=naturae vultus=in=orbe quem=dixere=chaos=rudis=
indigestaque=moles=nec=quicquam=nisi=pondus=iners=congestaque=
eodem=non=bene=iunctarum=discordia=semina=rerum nullus=adhuc=
mundo=praebebat=lumina=titan=nec=nova crescendo=reparabat=
cornua=phoebe=nec=circumfuso=pendebat=in=aere=tellus=ponderibus=
librata=suis=nec=bracchia=longo=margine=terrarum= porrexerat
e: 17.9%, 1-r: 93.1%, f: 1.9%

Bibliographie

BAVAUD F. (1998), *Modèles et données*, Paris, L'Harmattan.

BAVAUD F. (2000), «An Information Theoretical Approach to Factor Analysis», in *Proceedings of the 5th International Conference on the Statistical Analysis of Textual Data* (*JADT 2000*), pp.263-70.

BLOOMFIELD L. (1933), *Language*, Holt, New York

BRENT M.R. (1999), «An efficient, probabilistically sound algorithm for segmentation and word discovery», *Machine Learning Journal* 34, pp.71-106.

BRENT M.R. & CARTWRIGHT T.A. (1996), «Distributional regularity and phonotactics are useful for segmentation», *Cognition* 61, pp.93-125.

COVER T.M. & THOMAS J.A. (1991), *Elements of Information Theory*, Wiley, New York.

DELIGNE S. & BIMBOT F. (1995), «Language modelling by variable length sequences: Theoretical formulation and evaluation of multigrams», in *Proceedings of the International Conference on Speech and Signal Processing*, vol. 1, pp.169-72.

GAMMON E. (1969), «Quantitative approximations to the word», in *Papers presented to the International Conference on Computational Linguistics COLING-69*.

HARRIS Z.S. (1955a), «From phoneme to morpheme», *Language* 31, 190-222, réimprimé dans HARRIS Z.S. (1970), *Papers in Structural and Transformational Linguistics*, Dordrecht, D.Reidel, pp.32-67.

HARRIS Z.S. (1955b), *Methods in structural linguistics*, Chicago, University of Chicago.

HARRIS Z.S. (1967), «Morpheme Boundaries within Words: Report on a Computer Test», *Transformations and Discourse Analysis Papers* 31, réimprimé dans HARRIS Z.S. (1970), *Papers in Structural and Transformational Linguistics*, Dordrecht, D.Reidel, pp.68-87.

HARTLEY R.V.L. (1928), «Transmission of information», *Bell Systems Technical Journal* 7, 535-63.

HUFFMAN D.A. (1952), «A method for the construction of minimum-redundancy codes», in *Proceedings of the IRE.*

HUTCHENS J.L. & ALDER M.D. (1998), «Finding Structure via Compression», *Proceedings of the International Conference on Computational Natural Language Learning.*

KIT C. & WILKS Y. (1999), «Unsupervised Learning of Word Boundary with Description Length Gain», in *Proceedings CoNLL99 (Computational Natural Language Learning) ACL Workshop.*

MANNING C.D. & SCHÜTZE H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge, MIT Press.

DE MARCKEN C.G. (1996), «Linguistic structure as composition and perturbation», *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.*

MARKOV A.A. (1913), «An example of statistical inquiry on the text «Eugene Oneguine», illustrating tests on chains» (en Russe), *Bulletin de l'Académie Impériale de Sciences de Saint-Pétersbourg.*

MEL'ČUK I. (1997), *Cours de morphologie générale (théorique et descriptive) - vol.4. Cinquième partie, Signes morphologiques*, Presses de l'Université de Montréal, Montréal.

NEVILL-MANNING C.G. (1996), *Inferring sequential structure*, Doctoral thesis, University of Waikato, New Zealand.

OLIVIER D.C. (1968), *Stochastic Grammars and Language Acquisition Mechanisms*, Unpublished doctoral dissertation, Harvard University.

RISSANEN J. (1978), «Modelling by shortest data description», in *Automatica 14*, 465-71.

SAFFRAN J.R., NEWPORT E.L. & ASLIN R.N. (1996), «Word segmentation: The role of distributional cues», *Journal of Memory and Language 35*, 606-21.

SHANNON C.E (1948), «A Mathematical Theory of Communication», *Bell Systems Technical Journal* 27, 379-423.

WELSH. D. (1988), *Codes and cryptography*, Oxford University Press, Oxford.

WOLFF J.G. (1977), «The discovery of segments in natural language», *British Journal of Psychology* 68, 97-106.

XANTHOS A. (2000), «Entropizer 1.1: un outil informatique pour l'analyse séquentielle», in *Proceedings of the 5th International Conference on the Statistical Analysis of Textual Data* (*JADT 2000*), 357-64.