Approches et applications de la graphématique

Yannis Haralambous

Chapitre 7

Approches et applications de la graphématique

par Yannis Haralambous

Nous allons nous pencher sur un exemple d'interdisciplinarité par modèles interdisciplinaires, selon la typologie de Livet (cet ouvrage). Le sujet principal sera l'écriture : la linguistique construit un modèle structurel de l'acte de communication, la statistique un modèle mathématique de l'écriture informatisée, la stéganographie un modèle mathématique de la variation de l'écriture informatisée pour y insérer de l'information cachée, la biométrie un modèle physique de l'acte d'écrire sur clavier. Par la redécouverte de notions communes, toutes ces disciplines sont amenées à poser des nouvelles questions et, peut-être, de developer de nouveaux formalismes.

1.1. 7.1. Écriture et linguistique

L'écriture est une modalité de représentation du langage humain et de ce fait relève avant tout de la linguistique. Or la linguistique l'a longtemps reniée, discriminée, maudite : il suffit de citer le père-fondateur de la linguistique, Ferdinand de Saussure qui aurait dit dans ses cours :

> Langue et écriture sont deux systèmes de signes distincts ; l'unique raison d'être du second est de représenter le premier ; l'objet linguistique n'est pas défini par la combinaison du mot écrit et du mot parlé : ce dernier constitue à lui seul cet objet. Mais le mot écrit se mêle si intimement au mot parlé dont il est l'image, qu'il finit par usurper le rôle principal; on en vient à donner autant et plus d'importance à la représentation du signe vocal qu'à ce signe lui-même. C'est comme si l'on croyait que, pour connaître quelqu'un, il vaut mieux regarder sa photographie que son visage. [Saussure 1972, p. 45]

Cet extrait est doublement intéressant puisque d'une part il illustre la discrimination flagrante du langage écrit par Saussure (d'ailleurs il ne se donne même pas la peine d'indiquer que par «langue» il

entend la modalité orale de la langue) et d'autre part il est, sans que Saussure puisse l'imaginer à son époque, avant-coureur des problématiques actuelles de virtualité : à l'époque du numérique et de l'intelligence artificielle, y a-t-il toujours une différence entre une photographie et un visage ? Quoi qu'il en soit, l'attitude passionnelle de Saussure (ainsi que des néogrammairiens de Leipzig, qui l'ont précédé) est basée sur un principe que l'on peut appeler le «péché originel» de la linguistique : le *principe de la prépondérance du langage oral vis-à-vis du langage écrit*. Aristote déjà disait que

Ἔστι μὲν οὖν τὰ ἐν τῇ φωνῇ τῶν ἐν τῇ ψυχῇ παθημάτων σύμβολα, καὶ τὰ γραφόμενα τῶν ἐν τῇ φωνῇ [Aristote, 1936]

autrement dit : «les sons émis par la voix sont les symboles des états de l'âme, et les mots écrits les symboles des mots émis par la voix» (trad. J. Tricot, 1936), ce qui fait des «mots écrits» des citoyens de second rang du fait linguistique, puisqu'ils ne sont liés aux «états de l'âme» que par l'intermédiaire des «sons émis par la voix».

Il n'est donc pas étonnant que l'écriture soit délaissée par le *mainstream* des linguistes, même si de temps en temps quelques linguistes courageux (comme Vachek et Hořejsí de l'école de Prague) ont proclamé son importance égale avec celle de la parole.

Un autre facteur historique aggravant a été l'impossibilité de suivre le schéma classique de nommage des sous-disciplines de la linguistique. En effet, cette dernière a utilisé de manière très judicieuse les termes *phonétique* pour l'étude générale des sons émis par l'humain, et *phonologie* pour celle des classes de sons en tant qu'éléments distincts d'un système — appelés *phonèmes*. Le plus simple aurait été de suivre ce schéma et de définir la «graphétique», la «graphologie» et les «graphèmes». Or le premier terme n'a jamais été utilisé : il a fallu attendre les années 60 pour que quelqu'un se mette sérieusement à classer les «formes graphiques générales tracées par l'humain pour servir à la communication» (définition analogue à celle de la phonétique) ; le deuxième terme a été accaparé dès 1900 par la pseudo-science de la graphologie qui «affirme pouvoir déduire systématiquement des caractéristiques psychologiques de la personnalité d'un individu à partir de l'observation de son écriture manuscrite» (Wikipédia) ; seul le troisième terme est couramment utilisé (mais avec un grand nombre de définitions différentes, cf. [Pellat 1988]).

Il y a eu des tentatives d'invention de nouveaux termes : l'auteur utilise celui de *graphématique* en tant que pendant de la phonologie, d'autres ont proposé la «graphonomie», la «grammatologie» (ce terme, à l'origine introduit par Gelb [Gelb 1963] dans les années 60, est devenu célèbre par l'ouvrage homonyme de Derrida [Derrida 1967], qui relève plus de la philosophie que de la linguistique), et à un niveau supérieur : la «grapholinguistique» (d'après le terme allemand «Schriftlinguistik»), etc.

C'est dans les années 80 que les bases théoriques de la graphématique ont été établies. Jacques Anis [Anis 1988] énumère trois approches de cette discipline :

1. le *phonocentrisme*, représenté principalement par Saussure, qui considère que le langage écrit est subordonné à l'oral et le déforme;

2. le *phonographisme*, représenté par Vladimir Gak [Gak 1976] et Nina Catach [Catach 1986], qui considèrent que les langages oral et écrit sont des systèmes d'égale importance, mais que le deuxième est déterminé par le premier et ne peut exister sans lui;

3. l'*autonomisme*, représenté par Jacques Anis lui-même [Anis 1988], qui considère que l'écrit peut être étudié en parfaite indépendance de l'oral.

Ainsi, Catach [Catach 1978, p. 120] définit le *graphème* comme étant *la plus petit unité distinctive et/ou significative de la chaîne écrite, composée d'une lettre ou d'un groupe de lettres, ayant une référence phonique et/ou sémique dans la chaîne parlée*[1]. L'exemple qu'elle donne est le mot «pourchasser» qui comporte, selon elle, les huit graphèmes suivants : <p>, <ou>, <r>, <ch>, <a>, <ss>, <e>, <r>.

Elle classe les graphèmes en trois catégories :

1. les *phonogrammes* : graphèmes chargés de transcrire les phonèmes (ex. <g> dans *gare*) ;

2. les *morphogrammes* : notations de morphèmes, surtout situés, pour les renforcer, aux jointures des mots, maintenus graphiquement identiques qu'ils soient prononcés ou non (ex. : marques de féminin/masculin, singulier/pluriel, etc.) ;

3. les *logogrammes* : notations de mots, dans lesquels, à la limite, la «graphie» ne fait qu'un avec le mot, dont on ne peut la dissocier (ex. <à>/<a>, <coing>/<coin>, <fût>/<fut>, etc.). Pour Catach, la principale fonction des logogrammes est la distinction des homophones (même représentation phonétique, graphie et signification différentes) : «un logogramme n'est pas un *idéogramme* (c'est-à-dire la représentation d'une idée) puisque le son est noté, mais on y trouve plus que le son». Il existerait selon elle [Catach 1986, p. 268] des

[1] À noter que si on observe strictement cette définition, les lettres qui n'ont pas de référence phonique comme le <h> non aspiré en français, ne peuvent être des graphèmes. L'exemple du <h> posait problème déjà à Saussure [Saussure 1972, p. 53] qui disait : *l'*h *aspiré n'existe plus, à moins qu'on n'appelle de ce nom cette chose qui n'est pas un son, mais devant laquelle on ne fait ni liaison ni élision. C'est donc un cercle vicieux, et l'*h *n'est qu'un être fictif issu de l'écriture.*

> *lettres logogrammiques* qui ont comme fonction de distinguer le sens des mots et d'aider à leur connaissance immédiate visuelle. Mais plutôt que d'analyser dans ces cas chaque lettre comme un graphème de tel ou tel type, Catach propose comme alternative plus économique que l'on considère ces mots en tant que graphies globales — ainsi *coing* et *coin* sont pour elle des logogrammes.

Notons que Catach ne parle ni de ponctuation, ni de symboles tels que &, @, %, ni d'écritures autres que les alphabétiques.

Anis [Anis 1983, p. 33-34] propose une approche radicalement différente, mais complémentaire de celle de Catach. Pour lui :

- un *graphème* est une unité minimale de forme graphique de l'expression ;
- l'analyse des graphèmes se fait de manière analogue aux méthodes de la phonologie, c'est-à-dire par la méthode des paires minimales distinctives, méthode amplement mise en œuvre par les structuralistes ;
- si le mot existe bel et bien en écriture (Anis se limite aux écritures alphabétiques et *abjad*), il n'en est rien pour la syllabe : il n'existe aucun marqueur explicite des syllabes. Néanmoins on peut définir une *voyelle graphique*, «susceptible de former le noyau d'une syllabe» (par exemple, <a> et <y> étant des mots, ils sont par définition aussi des voyelles graphiques) et une *consonne graphique* qui «forme son satellite» (dans cette définition, Anis s'inspire des théories phonologiques, et notamment de l'approche de Malmberg [Malmberg 1971, p. 57]), il appelle les premières des *nodes* et les secondes des *sates*.

Cette courte introduction à l'histoire de la graphématique avait pour but de montrer au lecteur la difficulté d'étudier le langage écrit et les dilemmes qui en résultent : faut-il tenir compte du lien avec la phonologie ? Si non, et si on considère l'écrit sans la moindre influence phonologique — comme si on traitait l'écriture d'une langue éteinte ou d'une langue extraterrestre comme dans le film-fétiche des linguistes, *Premier contact* de Denis Villeneuve, 2016, inspiré de la nouvelle «L'histoire de ta vie» de Ted Chiang [Chiang 2017] —, y a-t-il moyen de retrouver les propriétés linguistiques de l'oral à travers l'écrit ?

Il s'avère que la réponse à cette question est affirmative, et la section suivante en constitue un brillant exemple.

7.2. La décomposition spectrale à la rescousse de la linguistique

Peut-on définir des nodes et des sates à la manière d'Anis, et retrouver deux classes de graphèmes correspondant aux voyelles et consonnes définies par les grammaires traditionnelles ?

Les linguistes canadiens Patricie Thaine et Gerald Penn de l'université de Toronto [Thaine & Penn 2017] y arrivent par une méthode ingénieuse. Ils utilisent la notion de *p-cadre*, définie comme suit :

Soit une chaîne de caractères informatiques[2] $C = c_1\, c_2\, \cdots c_n$, alors un *p*-cadre est une paire de lettres $(c_{i-1}\,, c_{i+1})$ (avec $1 < i < n$), que l'on note $c_{i-1}.\, c_{i+1}$ pour indiquer le fait qu'il «manque la lettre du milieu». Dire qu'«une lettre *a* participe à un *p*-cadre de *C*» signifie qu'il existe *b* et *c* tels que $bac \subset C$. Par exemple, les *p*-cadres du mot <dindon> sont {_.*i*, *d.n*, *i.d*, *n.o*, *o._*}, où _ symbolise les limites du mot. Notons maintenant *A* la matrice booléenne (de valeurs 0 et 1) dont les termes sont les cellules du tableau suivant :

	<d>	<i>	<n>	<o>
_.i	1	0	0	0
d.n	0	1	0	1
i.d	0	0	1	0
n.o	1	0	0	0
o._	0	0	1	0

où la colonne de la lettre <d> a deux entrées 1 puisque le grapheme <d> apparaît entre _ et <i>, et aussi entre <n> et <o>, et la ligne *d.n* a deux entrées 1 puisqu'on a un <din> et un <don>. Cette matrice représente les interactions de triplets de lettres, en se basant sur la lettre du milieu. Si sa largeur est limitée par le nombre de caractères étudiés (par exemple, 41 pour le français si on considère les 26 lettres standard, les lettres diacritées <à>, <â>, <ç>, <é>, <è>, <ê>, <ë>, <î>, <ï>, <ô>, <ù>, <û>, <ü>, <ÿ>, et le digraphe <œ>), sa hauteur peut être importante (jusqu'à $42 \times 42 = 1764$ si toutes les combinaisons de lettres apparaissent dans le corpus).

[2] Un caractère informatique est un élément de l'écriture défini par un consortium de fabriquants informatiques dans le cadre d'un *codage de caractères*, tel que Unicode [Haralambous 2007]. La majorité des caractères Unicode correspondent à des graphèmes, mais leur présence dans le codage ne résulte pas d'une approche scientifique par paires minimales comme dans l'approche d'Anis, mais de deux modélisations mécaniques historiques de l'écriture : de l'imprimerie et de la dactylographie.

En tant que matrice de nombres réels, A admet une *décomposition singulière*, c'est-à-dire qu'elle peut être écrite comme produits de matrices $U \times \Sigma \times V^*$, où Σ est diagonale, V^* est l'adjointe de V et U et V sont unitaires (c'est-à-dire que $U \times U^* = U^* \times U = \mathrm{Id}$, la matrice identité). On appelle les colonnes de V, les *vecteurs singuliers* de A. Notons σ_α et σ_β les deux plus grandes valeurs de Σ, où α et β sont les indices de σ_α et σ_β sur la diagonale de Σ. Prenons maintenant les colonnes d'indices α et β de V : ce sont des vecteurs de taille égale au nombre de lettres de notre système d'écriture, appelons-les x_* et y_*. Les i-èmes composantes x_i et y_i de ces vecteurs, correspondent à la i-ème lettre, on va les considérer comme *coordonnées* de la lettre dans le plan.

Ainsi, par exemple, pour l'exemple du mot <dindon>, on a la matrice A qui se décompose comme suit :

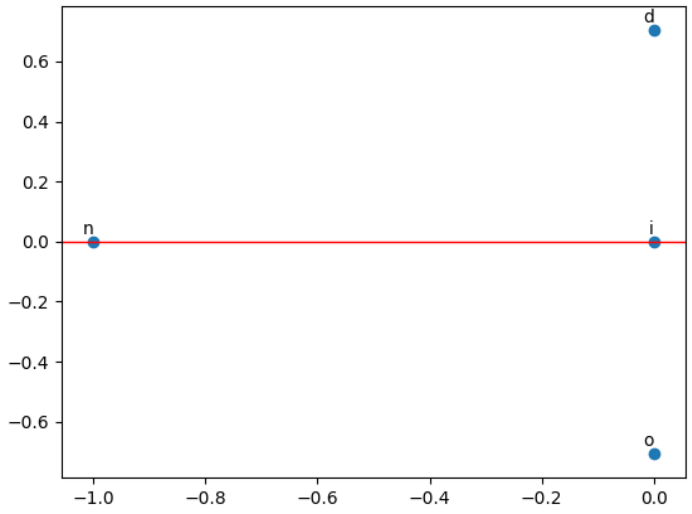
$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} = U\Sigma V^*,$$

$$\text{avec} \quad U = \begin{pmatrix} 0 & 0 & -0{,}707 & 0 & -0{,}707 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & -0{,}707 & 0 & 0{,}707 & 0 \\ 0 & 0 & -0{,}707 & 0 & 0{,}707 \\ 0 & -0{,}707 & 0 & -0{,}707 & 0 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 1{,}4142 & 0 & 0 & 0 \\ 0 & 1{,}4142 & 0 & 0 \\ 0 & 0 & 1{,}4142 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

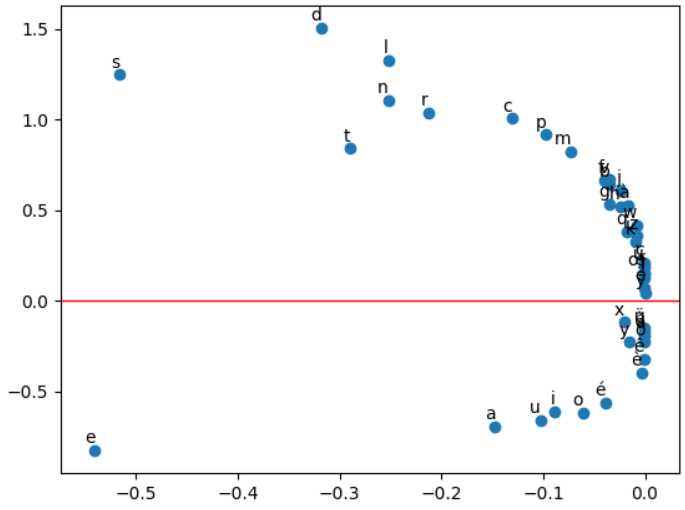
$$V = \begin{pmatrix} \mathbf{0} & \mathit{0{,}707} & 0 & 0{,}707 \\ \mathbf{0} & \mathit{0} & -1 & 0 \\ \mathbf{-1} & \mathit{0} & 0 & 0 \\ \mathbf{0} & \mathit{-0{,}707} & 0 & 0{,}707 \end{pmatrix},$$

où nous avons noté en gras le vecteur x_* et en italiques le vecteur y_*. Comme nous avons pris les lettres <d>, <i>, <n>, <o> par ordre alphabétique, la méthode nous fournit comme coordonnées de ces lettres les valeurs (0, 0,707), (0, 0), (–1, 0), (0, –0,707), que nous pouvons tracer sur un plan, comme suit :



Les auteurs de [Thaine & Penn 2017] stipulent que dans un diagramme obtenu de cette manière, les consonnes et les voyelles forment des clusters, disposés verticalement de part et d'autre de l'axe des *x*. Notre calcul a été trivial puisqu'on n'est parti que d'un seul mot, et pourtant nous avons déjà le résultat annoncé, puisque la lettre d se trouve au-dessus de l'axe des *x*, la lettre <o> au-dessous, et les deux autres lettres sur l'axe des *x*.

Pour avoir un résultat plus concluant pour la langue française, nous avons traité le corpus du Wikipédia français (datant du 1er mai 2017), comportant 947 783 813 mots (nous n'avons gardé que les 41 lettres citées ci-dessus). Après 2h30 de calcul, l'algorithme nous a sorti 1 718 *p*-cadres, et la décomposition spectrale de cette matrice nous a fourni ci-dessous :



où l'on voit très clairement[3] les consonnes et les voyelles séparées par l'axe des abscisses (à l'exception de la lettre <x> qui semble se trouver du côté des voyelles), et comme il a été prédit par [Thaine & Penn 2017], la semi-voyelle <y> est très proche de la ligne séparatrice entre voyelles et consonnes.

Pour conclure cette section, nous constatons que l'approche mathématique de la décomposition spectrale appliquée judicieusement aux graphèmes du français a permis de les classer en tant que représentants de voyelles ou de consonnes, avec une facilité et une précision surprenantes. L'approche spectrale est une généralisation de la méthode des paires minimales ; les valeurs positives de l'axe des *y* correspondent à la notion de node d'Annis et les valeurs négatives à celles de sate. On peut donc conclure que par la méthode de [Thaine & Penn 2017] on retrouve expérimentalement les nodes et les sates d'Anis, et ils correspondent assez naturellement aux voyelles et consonnes définies par les grammaires traditionnelles.

Dans la section suivante nous allons nous intéresser à une autre approche de l'écrit qui permet, elle aussi, de retrouver des caractéristiques linguistiques (dans ce cas, ce sera les limites syllabiques et morphémiques) par des opérations qui relèvent d'une autre discipline, la biométrie.

7.3. Application en biométrie

Les actes de *parole*, après avoir été étudiés par les linguistes, ont aussi intéressé les psychologues, les anthropologues, et plus tard les informaticiens spécialisés dans la reconnaissance de la parole. Les actes d'*écriture manuelle*, par contre, ont surtout été traités par une pseudo-science appelée «graphologie» (dont l'existence ne cesse de discréditer la graphématique). Mais récemment un autre type d'acte d'écriture a été étudié, celui où le scripteur utilise le clavier d'ordinateur comme outil d'écriture (cf. [Ballier 2019]).

En surveillant l'activité de saisie d'un sujet d'expérimentation, on récupère des données du type suivant (il s'agit de la phrase rituelle «Hello World !») :

Entry #		Key	Keycode	Loc.Press	Release
Num	1	? Shift	2	1114450735680	1114450736962
Num	2	H H	1	1114450735991	1114450736311
Num	3	e E	1	1114450737653	1114450738144
Num	4	l L	1	1114450738735	1114450739256
Num	5	l L	1	1114450739786	1114450740277
Num	6	o O	1	1114450740998	1114450741399
Num	7	Space	1	1114450742090	1114450742420
Num	8	? Shift	2	1114450743542	1114450745004
Num	9	W U	1	1114450743872	1114450744263

[3] Pour améliorer la lisibilité de cette image nous avons multiplié les y par $10^{0,75\times\log_{10}(\frac{M}{|y|})}$, où M est égal à $\max(y)$ ou à $|\min(y)|$, selon le signe de y. Cette function étant croissante, cela n'a pas affecté le rang des ordonnées.

```
Num   10  o   0       1   1114450745755   1114450746216
Num   11  r   R       1   1114450747017   1114450747437
Num   12  l   L       1   1114450748138   1114450748549
Num   13  d   D       1   1114450749310   1114450749771
Num   14  ?   Shift   2   1114450751373   1114450753776
Num   15  !   1       1   1114450752445   1114450752885
```

Dans cette figure tirée de [Villani 2007], l'on peut lire les temps d'enfoncement et de relâchement de la touche, à la milliseconde près[4]. Les lignes 1, 8 et 14 représentent la pression de la touche des majuscules, qui reste appuyée alors qu'on saisit la lettre en question. D'après [Villani 2007], à partir d'une saisie de 650 caractères, on peut identifier l'identité d'une personne (parmi 36) avec une précision de 99,4% (pour le même texte copié par tout le monde, ou 98,3% pour du texte librement choisi par chaque personne), ce qui constitue donc une méthode d'identification biométrique tout à fait raisonnable. Bien sûr, les performances baissent quand un individu saisie le texte sur des claviers différents, et 650 caractères demandent tout de même environ 1′30″ pour être saisis, ce qui est bien plus lent qu'un scan d'empreinte digitale ou de retine, mais cette méthode a l'avantage de pouvoir être appliquée avec très peu de moyens : un simple clavier d'ordinateur suffit.

En 2004, une équipe de linguistes allemands a étudié les IKI (*Interkey Interval* = délai entre deux touches) de textes allemands. Ils ont trouvé des résultats étonnants [Weingarten *et al.* 2004, p. 169] :

1. *segmentation sous-syllabique* : lorsqu'une syllabe dépassait les quatre caractères (ce qui arrive fréquemment en allemand où des mots à 10 lettres comme *quietschst* peuvent être *mono*syllabiques...) il y avait un IKI supplémentaire de 10 à 30 ms après la quatrième lettre, ce qui a été interprété par les psychologues comme un effet de segmentation de séquences motorisées, c'est-à-dire que le cerveau envoie des instructions aux doigts par séquence de 4 touches ;

2. *segmentation syllabique* : entre les syllabes intra-morphémiques (c'est-à-dire qui n'appartiennent pas à des morphèmes[5], différents), ils ont constaté des délais

[4] Les nombres des colonnes `Press` et `Release` représentent les millisecondes écoulées depuis le 1er janvier 1970, début de l'ère Unix. Ainsi `1114450752445` doit être lu comme 1 114 450 752 secondes et 445 msec, c'est-à-dire le 4 avril 2005, à 5h39 UTC et 445 millièmes de seconde.

[5] Un *morphème* est une suite de phonèmes ou de graphèmes qui constitue une unité minimale de sens, par exemple dans <brunes> on a trois morphèmes : le morphème lexical <brun>, le marqueur du féminin <e> et le marqueur du pluriel <s>. La langue allemande utilise abondamment la composition de morphèmes lexicaux (ainsi que grammaticaux), comme dans le mot <Wahrscheinlichkeitstheorie> (théorie des probabilités) comportant les morphèmes lexicaux <wahr> (vrai) et <schein> (apparence), le morphème grammatical d'adjectivisation <lich> (équivalent du suffixe français *-able*), le morphème grammatical de sub-

d'IKI supplémentaires de 40-70 ms. Ce délai n'est pas corrélé avec la fréquence du mot, et est le même aussi bien pour des mots existant que des mots fantaisistes ;

3. *segmentation morphémique* : entre les morphèmes ils ont constaté des délais supplémentaires de 100-150 ms, et ceux-ci sont effectivement corrélés avec la fréquence et la lexicalité du mot. Ce qui est normal, après tout si les morphèmes sont inventés, ils ne seront pas reconnus comme tels, et si les morphèmes existent mais pas le mot composé, alors le sujet prendra un «instant de réflexion» qui apparaîtra dans l'IKI intermorphémique.

Ainsi un mot comme <Flaschenöffner> (ouvre-bouteilles, constitué du morphème lexical <Flasche> (bouteille), du morphème grammatical <n> (marqueur du pluriel), du morphème lexical <öffn> (ouverture) et du morphème grammatical <er> (suffixe de l'agent)) sera saisi avec des délais d'IKI du type suivant :

Fla[40-70]sche[10-30]n[100-150]öf[40-70]fner.

On retrouve donc, par le biais de cette méthode, la structure morphologique du mot à travers le rythme de sa saisie, ce qui est révélateur de la manière dont le cerveau humain gère les données linguistiques, qu'elle soient liées aux graphèmes ou aux phonèmes. Ainsi, [Weingarten *et al.* 2004] mais aussi une équipe de chercheurs français [Bonnin *et al.* 2001], semblent pencher du côté de l'hypothèse autonomiste d'Anis qui, dans ce contexte, stipule que les représentations graphémiques sont indépendantes des représentations phonémiques.

Dans la section suivante nous allons envisager une autre approche de la graphématique : l'exploitation, à des fins cryptographiques, de la variation inhérente à une opération linguistique, la *transcription d'un système d'écriture dans un autre*. Cette variation a ses origines dans les rapports ambivalents et multiples de l'oral et de l'écrit, rapports qui se reflètent dans les différentes approches de l'écrit : phonocentriste, phonographique, autonomiste mais aussi étymologique/historique, et inspirée d'emprunts culturels.

7.4. Application en stéganographie

La *stéganographie* est une branche de la cryptographie qui traite des communications où la présence d'un contenu caché dans un message doit échapper aux observateurs tiers. Un exemple typique provenant de l'Antiquité romaine est celui des «nulles» :

---

stantivisation <keit> (équivalent du suffixe français *-ité*), le morphème grammatical <s> (génitif), le morphème lexical <theor>, et le morphème grammatical <ie> (terminaison du substantif).

> Le chiffrement par nulles est un genre de code camouflé. Cette méthode consiste à marquer d'un signe particulier certaines lettres d'un texte : seules ces quelques lettres sont porteuses de sens. Celles-ci sont dites «repérées». Le reste des lettres encadrant les lettres repérées sont appelées des nulles : elles sont dépourvues de signification. Leur mission est de tromper tout œil indiscret puisque le libellé de la missive ne sert qu'à masquer le texte réel. [Collard 2004]

7.4.1. Approche stéganographique du greeklish

Dans cette section nous allons décrire un travail commun avec Caroline Fontaine sur une méthode stéganographique basée sur les variations de transcription latine de texte grec. En effet, dans les réseaux sociaux de langue grecque il existe une pratique assez répandue qui est celle d'écrire le texte grec en caractères latins (communément appelé *greeklish*), pour de raisons de facilité (cela évite de changer de clavier virtuel) ou de style (effet de modernité ou d'adhésion à la culture occidentale). Or, cette pratique ne suit aucune norme spécifique et au moins cinq approches sont utilisées, souvent de manière simultanée :

1. l'approche inspirée de la philologie occidentale. Celle-ci considère que les lettres <θ>, <φ>, <χ> correspondent à des occlusives aspirées en grec ancien et donc que leur transcription est obtenue par la lettre représentant l'occlusive non aspirée (donc <t>, <p>, <c>) suivie du <h> représentant l'aspiration, on a donc <θ> → <th>, <φ> → <ph>, <χ> → <ch> ;

2. l'approche phonétique d'après le réseau de correspondances graphème-phonème de l'anglais ou de l'allemand : <φ> → <f>, <χ> → <h> (comme dans le nom de l'auteur), <η> → <i>, <ω> → <o>, <ψ> → <ps>, <οι> → <i>, etc. ;

3. des approches phonétiques *ad hoc*, comme par exemple <ξ> → <ks>, <χ> → <kh> (comme dans le cas de la transcription latine du nom russe «Khrouchtchev»), <γ> → <gh>, etc. ;

4. l'approche graphique, qui s'inspire des similitudes visuelles des graphèmes grec et latin : <ω> → <w>, <ς> → <c>, <η> → <h>, <χ> → <x>, <ν> → <v>, en allant parfois jusqu'à utiliser des chiffres pour certains graphèmes : <θ> → <8>, <ξ> → <3>, etc.

5. plus rarement, l'utilisation de la même touche du clavier : le fait que les lettres grecques <ς>, <θ>, <ξ>, <ψ>, <ω>, sont affectées aux mêmes touches du clavier grec que les lettres latines <w>, <u>, <j>, <c>, <v>, fonctionne comme un code et certains scripteurs utilisent les lettres latines pour représenter les lettres grecques affectées à la même touche, alors qu'elles n'ont ni lien

phonétique ni lien graphique entre elles. Alors <ψίθυρος> pourrait, par exemple, s'écrire <ciuyrow>. Cette approche n'arrive qu'épisodiquement et sans systématicité.

Pour illustrer nos propos, voici ce qu'a obtenu [Androutsopoulos 2001] après avoir demandé à 70 personnes d'écrire le mot <διεύθυνση> (= *adresse*) en *greeklish* : <diefthinsi> (12 fois), <diey8ynsh> (7 fois), <dieuthinsi> (7 fois), <dieftinsi>, <dieuthhnsh>, <dieuthunsi> (3~fois), <dieyuynsh>, <dieythinsi>, <dieuthynsi>, <diey8insi> (2 fois), et des *hapax* <diethinsh>, <dieu0unsh>, <dieu9unsh>, <diey0hnsh>, <dieu8uvsn>, <dieu8insi>, <dievthinsi>, <dieythynsi>, <diefhinsi>, <diey0yvsn>, <dieu0uvsn>, <dieuthunsh>, <dief8hnsh>.

La variété des approches illustre la créativité des scripteurs grecs mais surtout la multitude des approches de l'écriture : les approaches 2 et 3 reflètent le phonocentrisme qui règne, en tant qu'idéologie linguistique, sans partage en Grèce depuis le mouvement démotique du début du xx^{e} siècle, alors que l'approche 4 relève plutôt de l'autonomisme, puisque le graphème est considéré en tant que signe indépendant de toute correspondance phonétique. Enfin, l'approche 1 relève de l'influence de la vision que l'Occident a de la Grèce sur la langue et la culture du pays.

Sur un plan technique, ces approches sont souvent incompatibles entre elles, ce qui interdit certaines combinaisons de méthode au niveau de la même syllabe ou du même mot. Ainsi, le mot <φήμη> s'écrirait <fimi> selon l'approche 3, <phimi> selon un mélange des approaches 1 et 3, <fhmh> selon un mélange des approaches 2 et 4, mais la forme *<phhmh> qui résulterait de l'utilisation du <ph> pour <φ> (approche 1) et de <h> pour <η> (approche 4) est simplement illisible.

Notre méthode consiste à nous servir des choix d'approche de translittération en tant qu'information cachée dans le message. Aux incompatibilités logiques évoquées dans le paragraphe précédent s'ajoutent aussi des critères de vraisemblance : le fait d'utiliser deux méthodes différentes dans le même mot (comme par exemple en écrivant <tha8ela> le syntagme <θἄθελα> où le premier <θ> est transcrit selon l'approche 1 et le deuxième selon l'approche 4 : il n'y pas d'incompatibilité logique, mais il est invraisemblable que quelqu'un utilise les deux méthodes dans le même mot).

7.4.2. *Méthode stéganographique, évaluation*

Notre méthode consiste à choisir des méthodes de translittération différentes pour les mots d'un texte, selon l'information qui va être cachée dans le message. Imaginons, par exemple, que pour les graphèmes d'un mot il y ait deux méthodes de translittération possibles (ce qui est une sous-estimation, puisque certains graphèmes peuvent

être transcrits de pas moins de quatre manières différentes, comme le <χ> : <ch>, <kh>, <h>, <x>), ce mot porterait donc un bit d'information.

Pour tenter d'évaluer, ne serait-ce grossièrement, les performances de cette méthode, nous proposons les calculs suivants :

- admettons que seuls les 13 graphèmes suivants peuvent avoir des multiples translittérations : <γ>, <δ>, <η>, <θ>, <ν>, <ξ>, <ρ>, <ς>, <υ>, <φ>, <χ>, <ψ>, <ω> ;
- le corpus Wikipédia grec (datant du 1er septembre 2018) comporte 90 542 689 mots, dont 73 616 745 contiennent au moins un parmi les graphèmes cités ;
- le texte anglais de *Moby Dick* comporte 1 192 635 caractères, si on oublie les capitales et les chiffres et que l'on se restreint aux 26 lettres de l'alphabet et à 6 signes de ponctuation, ces caractères peuvent être codés sur 4 octets ;
- alors en divisant le nombre de mots du Wikipédia grec candidats à une variation de translittération par 4 fois le nombre de caractères de *Moby Dick*, on trouve que l'on peut cacher dans le premier une quantité d'information égale à 15 fois le texte du second, ce qui est énorme en soi ;
- et enfin, un autre exemple : le texte de la traduction en grec moderne de l'*Iliade* d'Homère, récupérable sur le site du Projet Gutenberg, comporte 111 583 mots dont 78 622 sont candidats à une variation de transcription, on peut donc y cacher un texte de 19 655 caractères latins, donc l'équivalent des neuf premières pages du texte que le lecteur a sous les yeux, espaces compris.

Bien sûr, cette évaluation ne tient compte que de la quantité d'information que l'on peut cacher à travers cette méthode, à cela il faut ajouter le degré de vraisemblance (ou de «naturalité») d'un texte qui alterne constamment entre différentes méthodes de translittération, et il se peut bien que notre méthode soit sous-optimale sous cet aspect.

Néanmoins elle illustre bien la multitude d'approches théoriques de l'écriture : le phonocentrisme, l'autonomisme, l'historicisme. Lorsqu'un scripteur choisit l'approche (2) de translittération du grec, il se place dans la continuité du courant phonocentriste qui résulte de la diglossie qui a régné dans l'espace linguistique grec entre le XVIIIe siècle et 1976, date d'instauration de la démotique comme langue officielle d'État. Lorsqu'il choisit l'approche (4) il adopte une approche autonomiste puisqu'il écarte la référence phonétique des lettres latines et s'en sert graphiquement. Lorsqu'il choisit l'approche (1), il se place

dans un contexte mondialiste où la longue tradition des hellénistes allemands, français et britanniques qui se reflète dans la transcription latine des mots grecs utilisée dans ces langues affecte sa vision des rapports entre les écritures. Pour résumer en un exemple : en écrivant le terme médical grec <ἄφθα> (devenu *aphthe* ou *aphte* en français et *Aphthe* en allemand) en tant que <af8a> il choisit de privilégier l'aspect visuel ; en écrivant <aftha> il choisit de donner la prononciation (pour le thêta il n'a pas le choix de faire autrement le seul graphème anglais correspondant à la fricative dentale sourde /θ/ est <th>, et ni le français ni l'allemand ne possèdent ce phonème) ; et enfin en écrivant <aphtha> il s'aligne avec la tradition helléniste et l'évolution graphique de ce mot en Occident.

7.5. Conclusion

La brève série d'exemples hétéroclites que nous venons de donner dans ce chapitre a comme point commun l'interaction entre disciplines face à un phénomène donné — en l'occurrence la modalité écrite du langage. On voit bien à travers les exemples que l'approche théorique (inspirée du structuralisme ambiant) qu'Anis introduisait en 1983 a pu être vérifiée mathématiquement en 2017 par la décomposition spectrale et que les notions de syllabe et de morphème ont été retracées biométriquement en 2006. Enfin, dans le dernier exemple, on a vu comment les différentes approches de l'écriture peuvent avoir des applications imprévues et inimaginées dans une discipline aussi éloignée du reste que la cryptographie.

Références

ANDROUTSOPOULOS J., «Ἀπὸ dieuthinsi σὲ diey8ynsh. Ὀρθογραφικὴ ποικιλότητα στὴν λατινικὴ μεταγραφὴ τῶν ἑλληνικῶν», AGOURAKI Y. *et al.*, éds. *Proceedings of the 4th International Conference on Greek Linguistics*, Nicosia, September 1999, Thessaloniki, University Studio Press, 2001.

ANIS J., «Pour une graphématique autonome», *Langue française*, vol. 59, p. 31-44, 1983.

ANIS J., *L'écriture, théories et descriptions*, vol. 10 de *Prisme Problématiques*, DeBoeck Université, Bruxelles, 1988.

ARITOTE, *De l'interprétation*, Vrin, Paris, 1936, traduction et notes par J. Tricot.

BALLIER N., PACQUETET E. & ARNOLDT T., «Investigating key-logs as time-stamped graphemics», *Proceedings of the /gʁafematik/ Conference*, June 14–15, 2018, Brest, 2019, (to appear).

BONIN P., PEEREMAN R. & FAYOL M., Do phonological codes constrain the selection of orthographic codes in written picture naming ? *Journal of Memory and Language*, vol. 45, p. 688-720, 2001.

CATACH N., *L'orthographe*, vol. 685 de *Que sais-je ?*, PUF, Paris, 1978.

CATACH N., *L'orthographe française, traité théorique et pratique*, Nathan, Paris, 1986.

CHIANG T., «L'histoire de ta vie», *La tour de Babylone*, Paris, *folio SF*, p. 137-211, 2017.

COLLARD B., «Les langages secrets dans l'Antiquité gréco-romaine», *Folia Electronica Classica*, vol. 8, 2004, http://bcs.fltr.ucl.ac.be/FE/08/stegano.htm.

DERRIDA J., *De la grammatologie*, Éditions de minuit, Paris, 1967.

GAK V.G., *L'orthographe du français*, Selaf, Paris, 1976.

GELB I., *A Study of Writing*, The University of Chicago Press, Chicago, 1963.

HARALAMBOUS Y., *Fonts & Encodings. From Advanced Typography to Unicode and Everything in Between*, O'Reilly, Sebastopol, CA, 2007.

MALMBERG B., *Les domaines de la phonétique*, PUF, Paris, 1971.

PELLAT J.-C., «Indépendance ou interaction de l'écrit et de l'oral ? Recensement critique des définitions du graphème», dans *Pour une théorie de la langue écrite*, Paris, Éditions du CNRS, p. 133-146, 1988.

SAUSSURE (DE) F., *Cours de linguistique générale*, Éditions Payot, Paris, 1972.

THAINE P. & PENN G., «Vowel and Consonant Classification through Spectral Decomposition», *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, ACL, 2017.

VILLANI M., *Keystroke Biometric Identification Studies on Long-text Input*, PhD thesis, New York, NY, USA, 2007.

WEINGARTEN R., NOTTBUSCH G. & WILL U., «Morphemes, syllables and graphemes in written word production», *Trends in linguistics studies and monographs*, vol. 157, p. 529-572, 2004.