*Cahiers de l'ILSL-45* (2016 : 19-44)

# Une analyse de la conjugaison française basée sur les données[1]

Gilles BOYÉ
Université Bordeaux-Montaigne & CNRS (F)
gilles.boye@u-bordeaux-montaigne.fr

## 1. Introduction

À quoi devrait ressembler une analyse de la conjugaison française basée sur les données ? D'abord, une telle analyse devrait probablement prendre en compte de larges quantités de données. Or, je ne dispose que de lexiques, c'est-à-dire d'une petite quantité de données, compilées à partir de données plus importantes collectées par d'autres personnes. Parlons donc en premier lieu de ce problème.

Il me manque un corpus, un vrai corpus duquel je pourrais tirer les formes conjuguées et les rentrer dans mon moteur pour effectuer un traitement informatique. Il me faudrait plus de matière, mais je n'en ai pas. L'autre élément nécessaire est de réduire au maximum le nombre des hypothèses. Je n'en utilise que deux. J'ai besoin de descriptions phonémiques des segments, parce que je travaille au niveau phonologique. Je n'utilise que des analogies simples entre les formes. Par exemple, si on a une représentation phonologique donnée dans une case du paradigme, et une autre représentation phonologique dans une autre case, comment passe-t-on de l'une à l'autre par analogie ? Ensuite, à partir des larges quantités de données – que je n'ai pas… –, je voudrais extraire des généralisations et générer des formes candidates, en partant d'une case par exemple, et tenter de

[1] *A data-driven analysis of French conjugation: driving around the curbs*. Transcription, traduction et adaptation par Guillaume Feigenwinter et Christian Surcouf.

remplir la totalité du paradigme d'un verbe, voire de tous les verbes. Ensuite, parce qu'il y a beaucoup de bruit, en raison du grand nombre d'analogies, j'essaie de faire un tri et peu à peu de trouver quels paradigmes peuvent être retenus, et lesquels devraient être éliminés. J'ai donc des problèmes avec les données et des problèmes avec les analyses.

## 2. Le problème avec les données

Commençons par les données. Dans une langue, il y a beaucoup de variation, ce qui s'avère difficile à gérer, et pas seulement au niveau de la prononciation, comme, par exemple, les réductions de mot (voir ERNESTUS, page 65). Les locuteurs ont par ailleurs tendance à avoir plusieurs formes conjuguées différentes pour une même case, et je ne sais pas quoi faire avec ce genre de problème. Quel que soit le niveau d'expertise des transcripteurs, des désaccords subsistent sur la phonologie. Alors que faire ? Encoder moins de choses, ne garder que les points sur lesquels tout le monde s'accorde, ou bien encoder plus et trouver un moyen de gérer les dissensions ? Il nous faut prendre une décision.

Quant au lexique, que dois-je mettre dedans ? Quels mots existent ? Quels mots n'existent pas ? Pour chaque mot, chacune de ses formes potentielles existe-t-elle, c'est-à-dire *concrètement* dans la langue orale ? En français, combien de fois avez-vous entendu le subjonctif imparfait à la deuxième personne du singulier d'un verbe ? De toute votre vie ? Probablement jamais ! Nous faisons l'hypothèse que la case existe, mais en réalité personne n'a jamais entendu ce qu'elle contient. Alors qu'est-ce qu'on fait ? On l'enlève ou on la laisse ? Est-ce que c'est une vraie case ? Voilà le problème avec les données en général, ensuite viennent ceux de l'analyse.

## 3. Les difficultés de l'analyse

Dès qu'on envisage la morphologie, se pose la question des phénomènes à analyser. C'est compliqué parce qu'en morphologie, bien des choses ne se déroulent pas comme prévu, comme par exemple la prononciation. Pour cette raison, des cadres théoriques très variés ont été proposés pour décrire des aspects très différents des données. Que devrait-on garder ou éliminer ? Toutes ces questions sont en fait élémentaires. Prenons l'exemple de la variation.

### 3.1. La question de la variation

Dois-je intégrer une forme comme « ils croivent », censée être socialement marquée ? Je ne dirais pas qu'elle est inacceptable parce qu'elle est marquée, d'ailleurs beaucoup de gens l'emploient. Dois-je faire comme certains et prétendre qu'une telle forme n'existe pas ou dois-je au contraire la prendre en compte puisque, effectivement, des locuteurs l'utilisent ? Et que faire de *frire* ? *Frire* a cette capacité intrigante d'être un verbe différent selon la zone de la francophonie dans laquelle on se trouve. Pour certains, c'est un verbe défectif, qui n'a que quelques cases remplies. D'autres pensent que c'est un verbe tout à fait normal, et que toutes ses cases sont remplies. Beaucoup de locuteurs pensent un peu des deux, et décrivent une défectivité plus ou moins importante. Devrais-je alors considérer *frire* comme complètement défectif, ou l'inverse puisque que certaines personnes en connaissent toutes les formes ? Et qu'en est-il de *chourave* ? Je ne sais pas ce qu'il en est en Suisse, mais en France, *chourave* est un mot étrange. Il ne varie jamais, il n'a qu'une seule forme pour moi, mais pour beaucoup de gens, c'est devenu *chouraver*, c'est-à-dire un verbe tout à fait régulier du premier groupe. Mais le verbe français que, moi, je connais se conjugue : « je chourave, tu chouraves, il chourave, nous… volons, vous… volez, ils chouravent » et « je l'ai chourave ». Vous voyez que le participe est aussi

« chourave », tout comme l'infinitif. C'est à peu près tout ce que nous avons : le participe passé, l'infinitif et quatre formes du présent de l'indicatif et du subjonctif, c'est tout. On a environ une demi-douzaine de verbes de la sorte, qui se terminent tous en *-ave* parce qu'ils proviennent de la même langue, et qu'ils ont gardé les mêmes propriétés. Cependant, avec le temps, ce verbe a évolué en un verbe normal. Devrais-je prendre le premier paradigme ou le nouveau paradigme ? Et de quelle population ? En fait, je ne sais pas.

Ces exemples illustrent des points de données spécifiques à un lexème, mais plus généralement se pose la question de savoir comment traiter l'intégralité du contenu d'un corpus. En effet, si j'ai un corpus global, alors il ne sera pas homogène. Est-ce représentatif de prendre des points de données d'individus, sachant que ces individus ont été choisis de manière à être représentatifs de leur population ? Ou devrais-je au contraire mélanger les données de plusieurs corpus et essayer de représenter l'entièreté des données, d'une manière ou d'une autre ? Là encore, je ne sais pas. Ma conférence n'est pas très informative…

Et jusqu'à quel point devrais-je manipuler les données ? Est-ce légitime ? Tout le monde le fait. On a parfois des données aberrantes, mais signifient-elles quelque chose ? Peut-on démontrer qu'elles indiquent quelque chose du signal, mais trop insignifiant pour qu'on en tienne compte ? Quelles sont exactement les données aberrantes ? La forme *croivent* est-elle une erreur ou un dialecte ? Comment décider ? Ce sont là des questions linguistiques générales. Abordons maintenant la question de la fréquence.

### 3.2. La question de la fréquence

À supposer que j'aie une grande quantité de données, j'ai l'avantage de pouvoir en tirer des fréquences. Je peux tenter d'avoir une couverture étendue, au lieu d'une couverture standard. Je peux essayer d'obtenir des fréquences orales, au

lieu de fréquences écrites, les seules dont je dispose pour le moment. Je n'ai aucune idée des fréquences orales : aucune donnée, rien, pourtant il est probable que la fréquence ait une influence sur le problème que je tente de résoudre. Maintenant, imaginons que j'aie toutes ces fréquences : comment savoir, quand on voit des formes homophones, laquelle est vraiment comptabilisée quand je compte les fréquences ? La forme *mange* serait-elle un subjonctif ? Une première, une troisième personne ? Est-ce important ? Si oui, quand ? Est-ce important au sein du même lexème, ou seulement entre différents lexèmes ? Est-ce important entre différentes catégories ? Je ne sais pas. Et comment traiter la défectivité ? Comment prendre en compte la fréquence d'une entité qui n'est jamais là ? Je m'intéresse plus particulièrement à la défectivité, parce que j'ai pendant longtemps travaillé sur ce problème. Au début, on ne me laissait même pas en avoir un modèle, on me disait : « Ça n'existe pas. », parce que ce n'est pas un point de donnée positif. À priori, on pourrait effectivement parvenir à une telle conclusion, mais l'observation des données invite au contraire à reconnaitre la défectivité. Par exemple, pour le verbe *clore*, bien qu'ils sachent que la forme est censée être *closons*, les locuteurs s'efforcent de l'éviter comme si leur vie en dépendait. Qu'en est-il maintenant des problèmes phonologiques.

### 3.3. La question de la phonologie

Imaginons que nous nous mettions d'accord sur un système phonémique. Il subsistera malgré tout le problème des voyelles moyennes en français. Alors, que fait-on des voyelles moyennes ? Dans bien des contextes en français, elles sont neutralisées, mais pas chez tous les locuteurs dans toutes les régions des pays francophones. À certains endroits, on distingue presque toutes les voyelles moyennes, alors qu'ailleurs on les confond presque toutes en une seule. Le

même type de problèmes se pose avec le schwa. Mais il y a plus encore, en français, en général, il n'y a pas de contraste entre un /ʁ/ et deux /ʁʁ/. Donc ce n'est pas comme si c'était un problème phonémique, ça ne vient pas des phonèmes eux-mêmes, mais de leur contraste. Il n'existe que dans trois lexèmes : *courir, acquérir, mourir*. Et on ne peut pas dire [ilkuʁa] (« Il *coura »), ce n'est pas du français. [ilkuʁa] n'est pas reconnaissable en tant que forme conjuguée de *courir*. On a la même chose avec *vingt-deux*, qui ne peut pas être prononcé [vɛ̃dø], qui correspondrait en l'occurrence à la séquence de *vingt* et *deux*, mais pas à *vingt-deux*. Quand je dis qu'on ne peut pas le dire, je parle depuis mon fauteuil, je ne suis pas allé vérifier dans un corpus. Je ne fais qu'écouter les gens. Pour *courir*, c'est évident. D'autres, comme [tʁije], soit « faire un tri », soit « émettre des trilles », ont exactement la même forme mais pas le même paradigme, parce qu'il y a neutralisation en français dans ce contexte. Il est impossible de distinguer la présence ou non d'une voyelle avant cette semi-consonne.

Il y a encore d'autres choses, comme quand vous faites un futur en français : le futur de *batter*[2] se prononce en général [batʁa], mais peut aussi être [batəʁa], alors que le futur de *battre* ne peut qu'être prononcé [batʁa] sans le schwa. Comment peut-on en rendre compte ? Devrait-on inclure le schwa dans tous les futurs de *batter*, en perdant le fait important que dans 99% des cas il n'y est pas ? Est-ce une séparation numérique entre sa présence et son absence ou a-t-on affaire à un phénomène graduel ? Et comment le représenter ? Au final, la seule chose dont je suis certain, c'est que j'aurai besoin d'un ensemble de traits. Il s'agit là de mon hypothèse de base, celle qui relève de ma responsabilité. Pour toutes les autres questions, je n'ai aucune idée.

[2] NdE : Frapper une balle à l'aide d'une batte.

### 3.4. Les problèmes avec le lexique

Venons-en maintenant à la question du lexique, qui soulève au moins un problème. Tous les lexiques ont été constitués automatiquement, et s'avèrent pour cette raison très cohérents. Comme on peut s'y attendre de tout procédé automatique, la moindre erreur est répercutée de manière systématique et constante à tout le corpus. Il est alors impossible de se fier à ce genre de lexique. Quant aux fréquences, elles proviennent de textes. Il n'y a pas de traitement spécifique de l'homophonie, ou de l'homographie, juste du texte, ce qui d'une manière générale n'a rien à voir avec de l'oral véritable.

J'ai deux sortes de lexiques. En premier lieu : les dictionnaires de paradigmes, très bien organisés, et s'intégrant parfaitement dans mon système parce qu'une forme donnée apparait dans une case spécifique, et qu'il n'y a qu'une forme par case. Mais les variantes et la surabondance ne sont en l'occurrence jamais prises en compte, ce qui, pour moi, constitue un problème, puisque je veux justement intégrer ces deux phénomènes. En second lieu, j'ai des dictionnaires de formes, qui ignorent l'existence des paradigmes. Ainsi perd-on le fait qu'*assoir* – verbe surabondant en français – a deux séries : *assois, assois, assoit, assoyons, assoyez, assoient,* et *asseye, asseyes, asseye, asseyons, asseyez, asseyent*. Avec un simple lexique de formes, il est impossible de savoir que ces deux séries vont ensemble, ce qui constitue une perte d'information.

Prenons le cas du verbe *courir*. Quelle est la forme qui apparait dans la case 'deuxième personne du pluriel du conditionnel présent' ? C'est un verbe très courant, que tout le monde connait et que tout le monde est convaincu de pouvoir conjuguer facilement. On hésiterait à priori entre

/kuʁje/ et /kuʁʁje/[3]. En français, on ne peut pas avoir deux /ʁ/ suivis d'un /j/, donc il faut changer quelque chose, par exemple enlever un /ʁ/. On peut également insérer un schwa, brisant ainsi le groupe consonantique. Certains locuteurs proposent même /kuʁiʁje/. Quoi qu'il en soit, tout le monde ressent la même chose et je suppose que tout le monde a un trou à cet endroit du paradigme. Ainsi évite-t-on d'utiliser cette forme. Ce n'est pas que nous sachions que les locuteurs évitent effectivement une telle forme, mais nous savons qu'ils devraient l'éviter parce qu'elle leur semble étrange. Les formes de ce genre ne sont pas à proprement parler problématiques, mais on a l'impression qu'elles ne devraient pas être dans cette case-là, qu'elles devraient en être expulsées, parce qu'elles ne se comportent pas comme les autres.

Si l'on évoque la défectivité, qui est mon domaine de recherche, on voit des choses comme *distraire*, dont les locuteurs ne savent pas former le passé simple. N'ayant jamais appris à le faire, ils n'en ont pas la moindre idée. Pour le pluriel de l'adjectif *nasal*, là on connait la réponse : c'est soit *nasals* /nazal/, soit *nasaux* /nazo/. Mais si vous enseignez la morphologie ou la phonologie comme moi, vous évitez ce mot au masculin pluriel, parce que soit vous dites /nazal/ et tout le monde rigole, soit vous dites /nazo/ et… tout le monde rigole ! C'est une autre sorte de défectivité, une défectivité apprise. Quand vous êtes au milieu d'une foule – comme je suis Français, on va dire que c'est une manifestation ou une grève, évènements courants en France – vous suivez les autres. Et si pour une raison ou une autre, les manifestants devant vous changent de route et se séparent, que faites-vous ? Continuez-vous tout droit ou suivez-vous tout simplement les gens ? Nous suivons normalement ceux qui

[3] Durant la conférence, Gilles BOYÉ a sollicité le public, et obtenu ces deux versions.

nous précèdent et c'est ce que nous avons tendance à faire avec *clore*. Nous remarquons que les gens ont tendance à l'éviter, nous ignorons pourquoi, mais c'est comme dans une grève ou une manifestation : si au lieu d'aller tout droit comme on s'y attend, la foule évite quelque chose, alors on choisit de faire pareil.

Comme je travaille sur la prédictibilité, je rencontre un autre problème. Par exemple, il est possible de prédire deux choses à partir de /ilɛ/. En fait, dès qu'on le dit à voix haute, le problème disparait : /il'ɛ/ (*il hait*) n'est effectivement pas la même chose que /ilɛ/ (*il est*). En français *oral*, il n'y pas d'ambigüité dans ce cas, et c'est seulement parce que les données dont je dispose n'incluent pas le fait que *hait* de *haïr* ne commence pas par une voyelle, au sens français du terme.

Voyons un autre problème courant : /ʒəsɥi/ *je suis* et /ʒəsɥi/ *je suis*. On sait que dans la réalité, les locuteurs ont différentes représentations pour ces deux-là, parce que l'énoncé « je suis une fille » n'est pas problématique. Si je dis [ʒəsɥiynfij], c'est évidement dans le sens de « je poursuis une fille », en revanche, si je dis [ʃɥiynfij], personne ne me croira, parce que [ʃɥi] est seulement une contraction possible pour le verbe *être*. Il y a donc, dans notre représentation, une différence concrète entre ces deux verbes dans cette case, invisible cependant dans les points de données à ma disposition.

Ensuite, il y a des choses identiques qui devraient effectivement l'être, parce que *assoir* reste le même mot, peu importe que nous ayons /aswa/ ou /asje/ ou /asɛj/, c'est d'une certaine manière toujours la même identité. Quant à *ficher*, c'est un verbe qui signifie « noter sur une fiche », et il n'a pas le même sens que *fiche*, employé dans la phrase *j'ai fichu mon vélo par terre*. Ils ont une grande partie de leur paradigme en commun, mais diffèrent en certains points. La plupart des Français les confondent et ne connaissent que la différence entre leurs participes passés.

### 3.5. La question des paradigmes

Venons-en aux paradigmes. Je pourrais faire des paradigmes syntactiques, comme tout le monde : une case par contexte syntactique avec des syncrétismes entre cases homophoniques. Ou alors je pourrais faire des paradigmes de formes. Et ensuite, il suffirait d'avoir un identifiant pour chaque forme, et des correspondances entre ces identifiants et les contextes syntactiques. Ainsi pour chaque forme, on aurait la liste des contextes syntaxiques où elle apparait. Ceci résoudrait le problème des homophonies et expliciterait la présence des syncrétismes, mais je perdrais alors l'uniformité de mes paradigmes, car les verbes auraient des paradigmes différents en fonction du nombre de formes qu'ils possèdent et des contextes syntactiques qu'elles occupent. En fait, mon problème est généralement de remplir le paradigme. Cela porte un nom : « le problème du remplissage des cases du paradigme »[4] (ACKERMAN, BLEVINS & MALOUF 2009).

Ces auteurs proposent une solution, en calculant l'entropie, définie comme la mesure du manque d'information à combler pour résoudre un problème. Selon eux, la morphologie flexionnelle est un problème à basse entropie, parce qu'en général il manque peu d'information pour remplir les cases des paradigmes. La flexion est effectivement un système très bien organisé à très faible entropie. Toutefois, leur solution présente un problème, que je ne discuterai pas ici. En effet, les auteurs ne proposent aucun modèle de flexion. Ils parlent juste d'entropie, mais jamais de flexion *réelle* : quelles sont les formes ? comment elles sont calculées ? Pour ACKERMAN, BLEVINS & MALOUF, il s'agit de remplir une case du paradigme à partir de l'information dont on dispose, mais ils ne précisent pas quelle

[4] NdE: « Paradigm Cell Filling Problem: What licences reliable inferences about the inflected (and derived) surface forms of lexical items » (voir ACKERMAN, BLEVINS & MALOUF 2009, 54).

est cette information. Il n'y a aucune mesure de la connaissance initiale. Pourtant dès qu'on travaille sur l'entropie, définie comme le manque de connaissance à combler pour résoudre un problème, une telle mesure devrait être obligatoire. Les auteurs prétendent inférer le contenu d'*une* case à partir d'une forme, mais en définitive, ils essaient d'inférer le contenu de *toutes* les cases à partir d'une seule forme. Même le fait qu'ils partent vraiment d'une forme n'est pas certain, parce qu'on n'a aucun contrôle sur la connaissance dont ils disposent au départ. En fait, il me semble qu'ils savent déjà presque tout du paradigme, et ils infèrent quel est le paradigme. Mon collègue, Olivier BONAMI a quant à lui mené de véritables recherches sur la manière dont on peut inférer le paradigme à partir d'une, deux, ou trois formes, en mesurant ce qu'il connait des formes initiales (BONAMI & BENIAMINE 2015). Bref, je vois là beaucoup de problèmes avec ces deux types de réponses à la question du remplissage des paradigmes et j'aimerais tenter d'en résoudre une partie.

### 3.5.1. La résolution des problèmes

Alors qu'en est-il de la variation ? Bien que je n'aie aucune idée de comment collecter les données, je sais en revanche comment les rentrer dans mon système. J'essaie donc de résoudre ce premier problème en intégrant non seulement des zéros et des uns, mais aussi des proportions, et des estimations. Pour la phonologie, j'essaie différents modèles, différents encodages, et j'analyse les résultats. À partir de là, je formule différentes hypothèses et tente de voir quelles sont les conséquences, et laquelle semble être la plus appropriée et adaptée aux données. Pour le lexique, je commence avec les lexiques que j'ai à ma disposition, bien qu'ils soient imparfaits, mais c'est tout ce que j'ai pour le moment. En ce qui concerne la morphologie, j'essaie de couvrir la supplétion, la défectivité et la surabondance. Quant

au cadre théorique, étant donné que je ne trouvais rien d'adapté, j'en ai proposé un nouveau, que j'ai nommé « distributions de paradigme ». Pour la morphologie des distributions de paradigme, je propose d'avoir des distributions pour les formes : plusieurs formes dans une case, mais pas n'importe quelle forme (par exemple, on aurait 10% d'une forme donnée, 20% d'une autre : c'est une distribution). Je procède de même pour les paradigmes tout en leur attribuant un indice de confiance. J'utilise également les réseaux « petit monde » (*small-world networks*), comme sur Facebook, où les amis de vos amis sont généralement vos amis. En d'autres termes, si vous êtes dans un paradigme, et que quelqu'un d'autre y est aussi, ses connaissances sont probablement dans le même paradigme que vous.

Je n'ai pas encore évoqué les problèmes de l'appariement un à un, généralement sous-jacent dans les modèles de la flexion. Étant donné l'absence d'analyse syntagmatique, je n'ai affaire qu'à des formes pas à des radicaux et des affixes donc le principe « un lexème, un radical » n'a pas de place ici. En ce qui concerne le principe « une case, une forme », je dois prendre en compte la surabondance et les variantes, fournies par mon échantillon lexical défini pour l'entrainement, et comme je n'ai pas encore de variante dans mon lexique, la variation ne peut être incluse. Quant au principe « un lexème, un paradigme », je sortirai plusieurs paradigmes si je constate qu'un lexème a plusieurs paradigmes possibles. En définitive, ce n'est pas un problème de remplissage de case, mais un problème de remplissage de paradigmes.

Venons-en aux implications des formes candidates. Depuis longtemps, nous avons des généralisations sur les implications. À ma connaissance, WURZEL (1984) est le premier à avoir proposé les conditions de structure paradigmatique du type : « Si ce mot a telle terminaison et est masculin dans cette case, alors il aura telle terminaison au féminin dans une autre case ». On parle alors d'implications

au sein du contenu. Pour les espaces thématiques, on aurait : « Si on a ce type de thème dans cette zone, alors toutes les racines dans cette zone sont identiques ». Par exemple, en français, le futur et le conditionnel ont toujours les mêmes thèmes. Même le verbe le plus irrégulier n'y échappe pas.

Abordons maintenant le modèle de l'» apprentissage par généralisation minimale » [5] que j'expliquerai plus loin. Présenté succinctement, il s'agit d'une manière de calculer les analogies entre des formes. La façon dont ALBRIGHT & HAYES procèdent est très spécifique : ils font la généralisation la plus petite possible pour chaque paire d'analogies qui subissent les mêmes transformations. Si un segment change d'une manière donnée dans deux contextes différents, ALBRIGHT & HAYES essaient de rassembler ces deux contextes et d'en proposer la plus petite généralisation possible. De telles implications ont été employées pour créer des modèles non pas de flexion, mais uniquement de prédictibilité.

### 3.5.2. Les « petits mondes »

Venons-en maintenant aux « petits mondes ». Les petits mondes sont utilisés en sociologie, parce que, apparemment, les réseaux sociologiques, comme les « réseaux sociaux », fonctionnent de cette manière : des individus connaissent d'autres individus qui eux-mêmes en connaissent d'autres, et les communautés sont en général des ensembles d'individus qui se connaissent, formant ce qu'on appelle une « clique ». Une clique est une partie d'un graphique, contenant des individus tous en relation mutuelle. Nous avons déjà utilisé les petits mondes dans notre laboratoire, pour des analyses de synonymie et de morphologie dérivationnelle, donc ça fait déjà partie de notre culture. Et en l'occurrence, c'est un très petit monde, puisque les paradigmes verbaux n'ont que quarante-huit cases en français, ce qui est très peu.

[5] NdE: « minimal generalization learner » (voir ALBRIGHT & HAYES 2002).

Cependant dès qu'on considère le lexique dans son entier, l'ensemble constitue un réseau immense. Chaque forme appartient à un petit monde, mais ils sont très nombreux dans le lexique.

L'avantage avec les petits mondes, c'est que les propositions de devenir amis proviennent d'individus probablement déjà connus. Si vous êtes relié à quatre personnes qui sont en relation avec un individu X, qui ne fait pas partie de votre réseau et si ces quatre personnes sont toutes reliées entre elles, alors la probabilité que vous connaissiez en fait cet individu X s'avère très élevée. Wiktionnaire utilise ce principe pour les synonymes. Si vous proposez un nouveau mot au Wiktionnaire, il vous dira en général : « pensez-vous que ce mot soit synonyme avec cette liste de mots ? » parce qu'il possède déjà un graphique des synonymes. Si vous suggérez un synonyme, il le prend et propose en retour tous les synonymes plausibles. Donc il y a des manières d'étendre des petits mondes et de créer plus de connexions par l'intermédiaire des voisins. Les petits mondes sont par ailleurs très stables. Si l'on enlève 10% des liens, et qu'on essaie d'étendre à nouveau les petits mondes, on verra qu'on retrouve probablement 99% de la configuration initiale.

Pour la flexion, un petit monde serait une clique. Dans un paradigme, les individus devraient tous être voisins. Je me sers de ce principe comme solution pour extraire des paradigmes : créer des petits mondes et ensuite en extraire des cliques. Voici ma procédure : j'analyse le lexique et j'essaie de créer toutes les analogies possibles entre chaque forme d'un même lexème en cherchant à faire des généralisations de manière minimale sur tout le lexique. Pas minimale, dans le sens d'ALBRIGHT, qui s'efforce d'avoir une généralisation minimale très fine. Moi, je garde seulement la généralisation la plus importante, celle qui a le contexte le plus vaste. Un contexte aussi généralisé que possible, pas

quelque chose de trop fin. À partir de ces analogies, je produis chaque forme, en partant de toutes les représentations lexicales d'un verbe. En définitive, pour chaque case du paradigme, je teste toutes les analogies qui fonctionnent. J'obtiens alors un réseau de grande taille, dont j'essaie d'extraire des cliques pour voir si je peux y trouver un paradigme utile. J'emploie le modèle d'apprentissage par généralisation minimale – ou plutôt une ré-implémentation, mais c'est presque pareil – et j'extrais toutes les analogies. Un tel modèle ne peut extraire que des affixes, ce qui s'avère suffisant pour le français. Je ne conserve que le contexte général, sans indice de confiance, parce que ceux fournis par le modèle d'apprentissage par généralisation minimale sont inutiles pour ma recherche.

Par exemple un tel modèle permet d'extraire la règle : « si on part d'un imparfait et qu'on passe au présent, la règle la plus générale est d'enlever le [ɛ] à la fin du mot ». Cette règle fonctionne pour 5237 verbes sur 6440, ce qui constitue une bonne généralisation, mais j'ai peut-être une vingtaine d'analogies différentes pour ça. Sur tout le lexique, en fusionnant l'ensemble, j'obtiens environ 1600 analogies de case à case, de mot à mot. Avec ces analogies, on n'utilise pas directement le modèle d'apprentissage par généralisation minimale, mais on essaie d'appliquer à nouveau les analogies aux lexèmes, et cela nous donne une distribution des possibilités qui existent réellement dans le lexique. Donc pour chaque forme, on regarde toutes les analogies qu'elle pourrait engendrer et on compte chaque fois que l'analogie est correcte. On parvient alors à une distribution qui couvre le lexique entier, de la manière dont cette ambigüité est effectivement réalisée.

Prenons un exemple. Disons que vous regardez les classes de la figure 1 (page suivante). Si vous regardez la classe 29, c'est censé être ambigu parce que les individus dans cette classe pourraient passer par ces deux règles. Mais en réalité,

aucun ne passe par cette règle-ci, la totalité des cas passent par cette règle-là. Ce n'est pas toujours le cas, comme on peut le voir avec la classe 27, on a plutôt une distribution, et l'output qu'on obtient suit la distribution du lexique d'entrainement (82% de /ɛ/ → [] et 18% de /tɛ/ → [] dans les contextes correspondants à ces deux transformations).

```
class 27 ( syportE ~ syport ):  231 members  points
E --> [] / X[p,t,k,b,d,g,f,s,S,v,z,Z,m,n,J,j,l,r,w,H,i,y,E,6,u,o,ê,û,ô] ___ # : 189—81.82% (supporter, etc.)
tE --> [] / X[p,t,b,d,f,s,v,z,m,n,r,E,6,a,o,ê,û,â,ô][r,E,a,ê,â] ___ # : 42—18.18% (sortir, etc.)
class 28 ( fryskE ~ frysk ):  4027 members  points
E --> [] / X[p,t,k,b,d,g,f,s,S,v,z,Z,m,n,J,j,l,r,w,H,i,y,E,6,u,o,ê,û,ô] ___ # : 4027—100.00% (frusquer, etc.)
local conditional entropy: -0.0
class 29 ( r6d6vE ~ r6dwa ):  2 members  points
E --> [] / X[p,t,k,b,d,g,f,s,S,v,z,Z,m,n,J,j,l,r,w,H,i,y,E,6,u,o,ê,û,ô] ___ # : 0—0.00%
6vE --> wa / X[t,d,s,z] ___ # : 2—100.00% (redevoir, etc.)
```

**Figure 1 – Distribution entre deux analogies**

On commence avec une représentation du paradigme, qui pourrait être un paradigme entier, une seule forme ou encore plusieurs formes dans une case, qui suivent une distribution : toutes les configurations sont imaginables. On a soit des variantes dans une case, avec des pondérations différentes, ou diverses formes du paradigme et on établit des inférences à partir de l'ensemble. On n'a par conséquent aucune des restrictions qu'on aurait eues si on partait du principe qu'il n'y a qu'une forme par case. Avec des formes cohérentes dans différentes cases, on peut commencer avec ce qu'on veut et voir où ça mène. À titre d'exemple simple, prenons le cas où l'on a une forme par case. Il nous faut développer une paire de paradigmes, parce que la première fois, on a seulement ce que le premier individu indique comme étant ses amis, mais on veut également savoir qui sont les amis de ses amis. Il faut donc développer deux paradigmes. La première et la deuxième fois suivent le même procédé. On prend toutes les formes de toutes les cases et on calcule les analogies. Dans la

figure 2 ci-dessous, les analogies n'ont pas toutes été calculées.

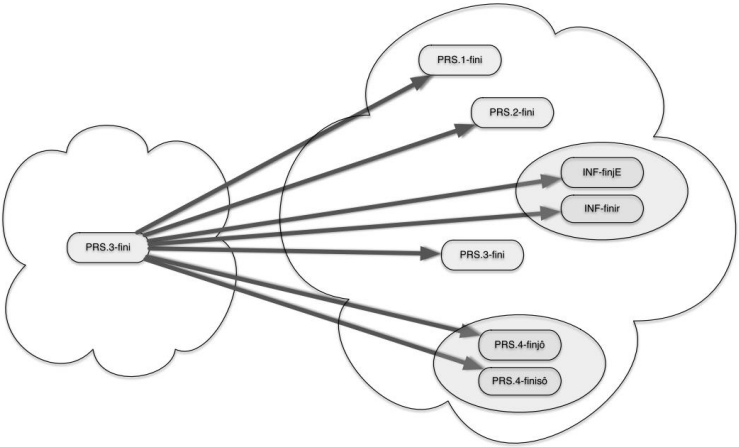


**Figure 2 – L'expansion d'un paradigme : première étape**

On obtient l'expansion des paradigmes en procédant de la même manière une seconde fois.

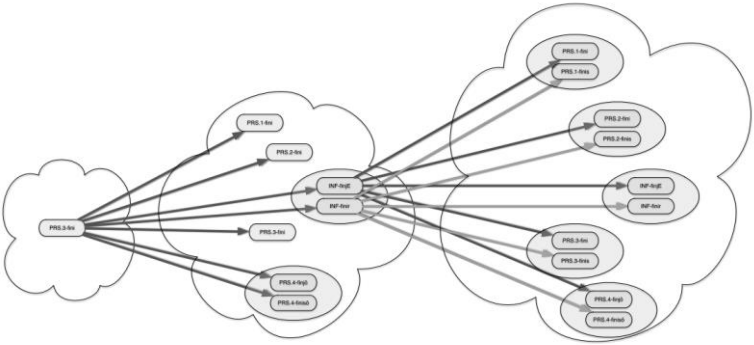

**Figure 3 – L'expansion d'un paradigme**

Dans la figure 3, la partie générée à la seconde étape et ses liens avec le point précédent sont importants pour collecter les paradigmes, et trouver les cliques. Pour ce faire, au sein

des deux paradigmes obtenus, on examine tous les points connectés, qui forment donc des cliques, qui sont ensuite extraites et devraient constituer le paradigme.

Prenons un exemple (voir figure 4), en démarrant avec la forme /kas/ pour l'impératif 2SG, au terme des deux étapes, on obtient un graphe qui ne contient qu'une seule clique, et donc un seul paradigme correspondant directement à celui de *casser*, rien de compliqué. Il n'existe aucun risque de confusion avec un autre verbe. Cette clique ne pose aucun problème.

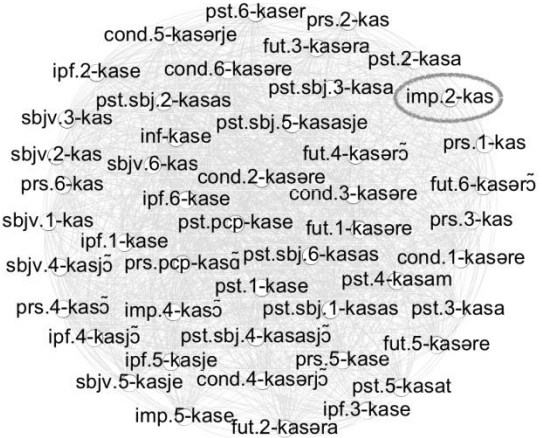


**Figure 4 – Graphe à partir de 'imp.2-kas' : une seule clique**

Maintenant, si on part de /kas/ non plus à l'impératif 2SG mais au présent 3PL, c'est une autre histoire, parce que 'prs.6-kas' produit plusieurs cliques (voir figure 5, page suivante). Il existe certes une clique complète de 48 formes qui constitue le paradigme du verbe *casser*, mais d'autres formes apparaissent qui n'appartiennent pas au paradigme de *casser*, et qui ont cependant été produites parce que ces analogies pourraient théoriquement convenir. Apparaissent notamment les participes passés *cassi et cassu*.

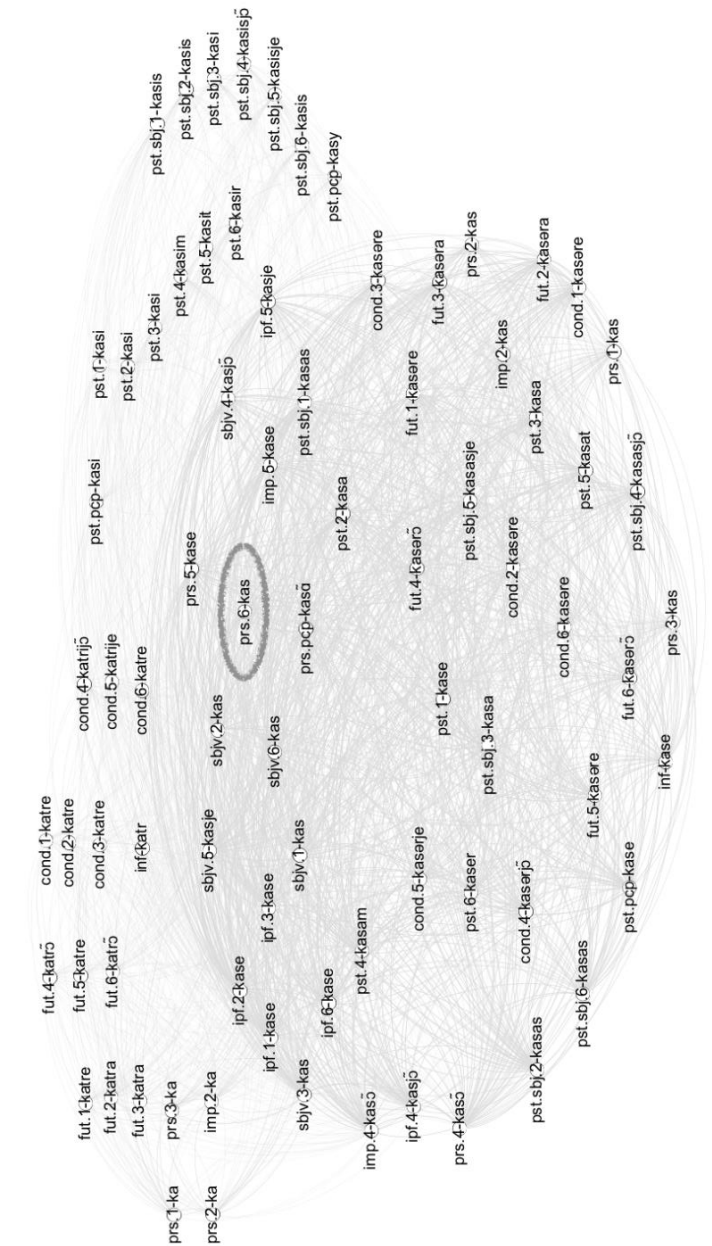


**Figure 5 – Graphe à partir de 'prs.6-kas' : plusieurs cliques**

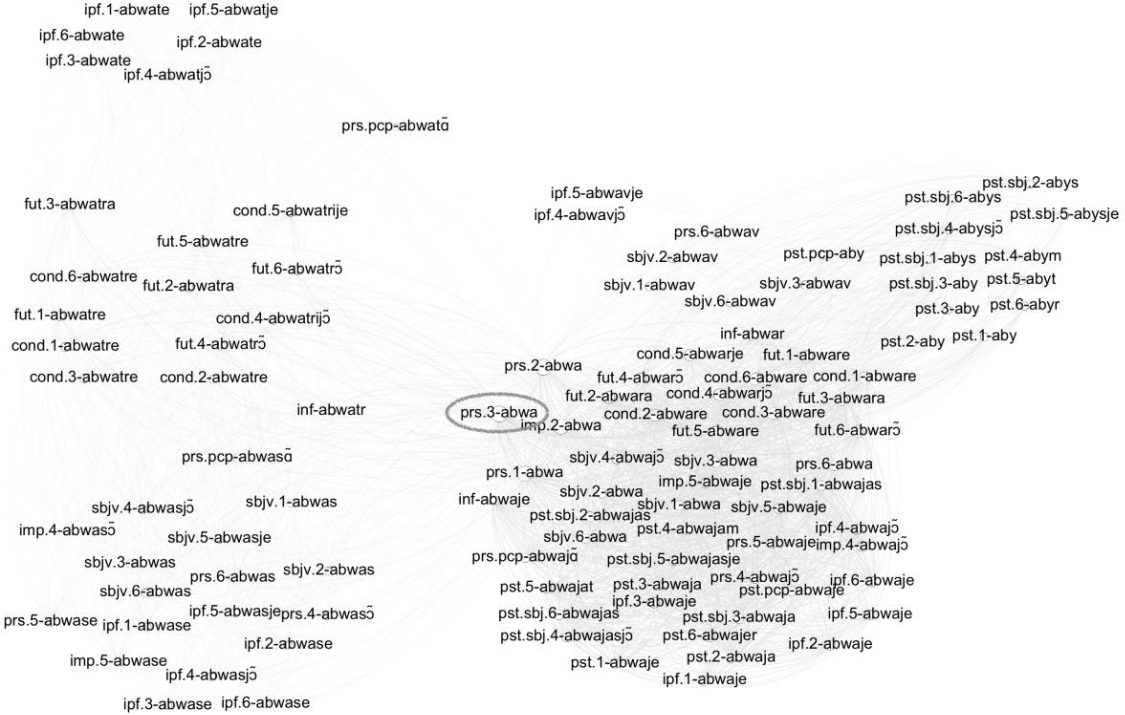


**Figure 6 – Graphe à partir de 'prs.3-abwa'**

Si nous observons maintenant le graphe de la figure 6 (page précédente), en partant de /abwa/ (*il aboie*), on obtient effectivement le verbe *aboyer* (en bas à droite sur la figure). Cependant on a également d'autres éléments étranges. Certains relèvent de la conjugaison normale du verbe *aboyer*, d'autres pas du tout. Ainsi en est-il de la partie en haut à droite qui donne l'impression qu'on a appliqué le paradigme de *boire* à *aboyer*. Mais si on extrait les cliques, on a bien le paradigme d'*aboyer* avec ses quarante-huit formes. Toutefois, la seconde clique sur le graphique est la suivante (tableau 1). Elle ne contient que trente-cinq membres, et s'avère par conséquent incomplète.

**aboyer-PRS3 35 0.541714285714**

| temps | 1SG | 2SG | 3SG | 1PL | 2PL | 3PL |
|---|---|---|---|---|---|---|
| present | abwa | abwa | abwa | | | abwav |
| imperfective | | | | | | |
| past | aby | aby | aby | abym | abyt | abyr |
| future | abware | abwara | abwara | abwarɔ̃ | abware | abwarɔ̃ |
| present subj. | abwav | abwav | abwav | | | abwav |
| imperfective subj. | abys | abys | aby | abysjɔ̃ | abysje | abys |
| conditional | abware | abware | abware | abwarjɔ̃ | abwarje | abware |
| imperative | | abwa | | | | |

| | |
|---|---|
| inf | abwar |
| pst.pcp | aby |

**Tableau 1 – La seconde clique de /abwaje/**

Le nombre indiqué en haut est l'indice de confiance pour cette clique, qui en l'occurrence s'avère plutôt bonne, avec des liens très forts, puisque /abwa/ (*aboie*) et /bwa/ *boit* sont très proches. Cependant, le réseau ne nous donne pas les quarante-huit formes attendues.

Pour conclure, le procédé proposé ici permettrait de modéliser la flexion sur la base d'analogies entre formes, la surgénération serait tempérée par la recherche de petits

mondes entièrement connectés qui formeraient les paradigmes flexionnels, le meilleur paradigme flexionnel correspondant à une clique couvrant toutes les cases du paradigme flexionnel.

## Questions

Légende : « Q » pour « Question », « GB » pour « Gilles BOYÉ »

**Q** : J'essaie de comprendre quelle est la nature des données que vous entrez dans votre modèle, et je comprends – ou du moins je crois comprendre – que vous avez des formes qui sont déjà complètement analysées, dans le sens où vous savez que c'est par exemple une troisième personne du singulier…

**GB** : Oui, elles sont complètement analysées, puisqu'elles sont classées dans telle ou telle case.

**Q** : Mais dans ce cas, d'où provient /aby/ (voir tableau 1, page précédente), je ne comprends pas.

**GB** : En fait, la première chose que nous faisons est d'analyser le lexique pour avoir un point de départ et pouvoir dire : « Ceux-ci, on les connait. » Nous avons la base, c'est tout ce que nous savons à propos des formes, mais nous n'avons pas les formes, tout ce que nous gardons, ce sont les analogies. Vous pouvez alimenter le programme avec une ou plusieurs formes à la fois, et c'est tout ce qu'il connait. Seul le programme connait toutes les paires entre les 6048 cases – mais il ne sait rien à propos d'une forme en particulier, il ne connait pas le lexique, seulement les analogies qu'il en a extraites. Donc quand il tombe sur quelque chose comme *avoir*, une analogie simple est de dire que ça ressemble à *voir*. L'analogie commence en appliquant le paradigme de *voir* à *avoir*. Mais à la deuxième étape, les analogies vont étendre cet embryon de paradigme sans pour autant parvenir à en étendre la portée sur l'ensemble des cases. De ce fait, cette famille d'analogies ne forme pas une clique complète. Voilà comment nous avons extrait une clique de seulement trente-cinq éléments, et une autre de quarante-huit, parce que des liens manquent. Seules les règles sont reconstruites, dans le cas de la clique complète. Le programme ne connait pas le lexique, mais il sait des choses à propos de ce lexique.

**Q** : Il sait des choses telles que « si un verbe se termine en /vwaʁ/, alors il pourrait bien donner /vy/ au participe passé » ?

**GB** : Il a des informations sur les cases. Donc si dans une case, on a une terminaison en /vwaʁ/, il se peut qu'une autre case contienne une terminaison en /vy/. Ça, il le sait grâce à la généralisation de cette alternance entre /wa/ et /y/, à partir de verbes concrets.

**Q** : Vous n'avez qu'une forme en entrée ou bien…

**GB** : Des paires, nous n'avons que des paires !

**Q** : Vous avez besoin de deux formes, de deux formes conjuguées pour générer le paradigme ?

**GB** : J'ai essayé ça, ce n'est pas un problème. Il est possible de partir d'un paradigme rempli, tout comme il est possible de partir d'une forme seule, ou de n'importe quelle configuration entre ces deux extrêmes.

**Q** : Qu'est-ce qui se passe si vous avez deux formes divergentes ? Par exemple si vous prenez une forme d'un de vos paradigmes et l'autre forme à partir de…

**GB** : J'ai essayé avec *assoir*, et vous arrivez à deux cliques.

**Q** : Donc par exemple, une pour *assieds* et une autre pour *assois* ?

**GB** : Oui, et elles sont distinctes. C'est parce que pour commencer, mon lexique n'a pas de surabondance. J'imagine que si j'incluais la surabondance dans mon lexique au départ, j'obtiendrais de la surabondance dans mes résultats, avec probablement des cliques plus grandes que quarante-huit formes.

**Q** : Quelle pourrait-être la réalité psycholinguistique de ce modèle ? Y en a-t-il une, est-ce que vous la cherchez ?

**GB** : Deux choses à ce propos : mes précédents travaux ont été critiqués sur ce point, et ma défense, auparavant, aurait été de dire « comment est-ce que vous rendez compte de la grammaire des locuteurs adultes ? » et je n'ai pas la moindre idée de comment on peut prendre une grammaire d'enfant et l'amener linéairement à une grammaire d'adulte. Alors tout le monde me disait que ça n'était pas possible, parce que linéairement, on ne peut pas aller d'un point donné à un autre point. À mon avis, en tout cas, il n'y a

pas de raison que ce soit linéaire. Des réorganisations se produisent en permanence, et la progression en U a quelque chose à voir là-dedans. Deuxièmement : cette fois-ci, j'ai pris quelques précautions, et en fait je pense que si vous entrez les données que vous voulez dans le système, vous n'avez plus qu'à savoir qu'il faut associer tel point de données avec tel autre. Vous devez savoir que c'est le même lexème, à un moment, mais vous pourriez trouver les analogies de toute façon, comme vous le voulez, et contrôler le réseau à chaque étape. Donc cela pourrait être linéaire, de cette manière.

**Q** : Et les fréquences ?

**GB** : J'ai fait une longue liste de problèmes, et la fréquence en fait partie, je ne veux pas m'y attaquer, parce que je n'ai pas les données nécessaires. Si c'était le cas, je proposerais d'utiliser des distributions, dans les cases et dans les paradigmes, mais je n'ai rien que je puisse employer.

**Q** : Quand même, dans le *Français fondamental*[6], vous savez que *bois* apparait là comme dans *je bois*, *tu bois*.

**GB** : Non, c'est une vraie question : de quelle fréquence parle-t-on ? Parce que j'en ai besoin, il me faut une idée de la distribution. À défaut, je n'ai rien. Je sais juste que *bois*, en tant que lexème, est fréquent. Mais il pourrait être pertinent de savoir qu'une forme donnée est fréquente, qu'une autre ne l'est pas, mais je ne peux pas faire de généralisation sans données concrètes.

**Q** : J'ai pensé à NEW, BRYSBAERT, VERONIS & PALLIER (2007) qui s'occupent d'analyser des sous-titres de films en français. Est-ce que ça aiderait ?

**GB** : Non.

**Q** : Et pourquoi pas ? C'est au moins un reflet écrit de l'oral, donc vous pouvez en savoir d'avantage sur les différentes formes.

**GB** : Bon, j'ai utilisé les fréquences des sous-titres. Je les ai regardées, mais il n'y a pas de futurs. Pourtant, j'entends tout le temps des futurs ! Si je veux intégrer les fréquences, il me faudrait

[6] NdE : le locuteur fait allusion à l'ouvrage de GOUGENHEIM, MICHÉA, RIVENC & SAUVAGEOT (1964).

des données concrètes, provenant d'expériences concrètes de locuteurs concrets. Les sous-titres, c'est loin, loin, très loin de remplir ces conditions. J'adorerais regarder d'autres corpus. Je sais juste que les sous-titres, ça ne va pas suffire.

**Q** : D'accord.

**GB** : Mais c'est effectivement une des choses que je dois dire : il me faut des données concernant les fréquences. Est-ce que les lexèmes les plus fréquents ont plus d'influence, ou au contraire moins d'influence sur le système ?

**Q** : Mais c'est intéressant, non ? Le fait qu'on ne puisse pas faire de prédiction est une bonne raison de s'y intéresser.

**GB** : En fait, depuis le départ, ma méthodologie a été de travailler avec des psycholinguistes, parce que utiliser les formes du *Bescherelle*, tout le monde peut le faire, de toutes les manières que vous voulez : il n'y a pas de restriction dans l'analyse. La seule contrainte intéressante pour l'analyse, c'est « Comment ça fonctionne dans le cerveau ? », « Comment est-ce que les gens produisent une conjugaison ? », « Comment ça fonctionne ? ». Ce n'est pas intéressant de connaitre *je crois, tu crois, il croit*, etc. Vous pouvez apprendre le *Bescherelle* par cœur, mais d'après ma propre expérience, j'en connais un rayon sur la conjugaison espagnole et italienne, et pourtant je ne peux pas utiliser des verbes correctement ni en espagnol, ni en italien. Je sais tout ce qu'il y a à savoir sur eux, et comment les former, si on me laissait prononcer un mot par minute… mais ça n'a rien à voir avec la connaissance des vrais locuteurs. Donc il me faut cette information psycholinguistique, tout comme il me faut les fréquences réelles, mais comme je l'ai déjà dit : je ne sais pas comment faire pour les obtenir.

## Références

ACKERMAN Farrell, BLEVINS James P. & MALOUF Robert (2009). Parts and Wholes: Implicative Patterns in Inflectional Paradigms. In BLEVINS James P. & BLEVINS Juliette (Eds), *Analogy in Grammar: Form and Acquisition*. Oxford: Oxford University Press, 54-82.

ALBRIGHT Adam & HAYES Bruce (2002). Modeling English Past Tense Intuitions with Minimal Generalization. In *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*, MAXWELL Michael (Ed.), Philadelphia: ACL, 58-69.

BONAMI Olivier & BENIAMINE Sarah (2015). Implicative Structure and Joint Predictiveness. In PIRRELLI Vito, MARZI Claudia, FERRO Marcello (Eds), *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*, Pisa, March 30-April 1, 2015: http://ceur-ws.org.

GOUGENHEIM Georges, MICHÉA René, RIVENC Paul & SAUVAGEOT Aurélien (1964), *L'Élaboration du Français Fondamental : Étude sur l'Établissement d'un Vocabulaire et d'une Grammaire de Base*, Paris : Didier.

NEW Boris, BRYSBAERT Marc, VERONIS Jean & PALLIER Christophe (2007). The Use of film Subtitles to Estimate Word Frequencies, *Applied Psycholinguistics 28*, 661-677.

WURZEL Wolfgang U. (1984). Flexionsmorphologie und Natürlichkeit. Ein Beitrag zur morphologischen Theoriebildung, *Studia Grammatica 21*, Berlin: Akademie-Verlag.