20 ans après : une transition écologique pour les espaces thématiques

Gilles Boyé

20 ans après : une transition écologique pour les espaces thématiques

Gilles Boyé

Université Bordeaux Montaigne & CLLE Montaigne

Gilles.Boye@u-bordeaux-montaigne.fr

Résumé. Pendant une dizaine d'années, la morphologie thématique a connu un succès pour les analyses flexionnelles et dérivationnelles en français (Bonami & Boyé, 2003, 2005, 2007, 2014 ; Bonami, Boyé, Giraudo & Voga, 2008 ; Bonami, Boyé & Kerleroux, 2009 : Plénat 2008 ; Roché, 2010 ; Roché & Plénat, 2014 ; Tribout, 2010) et dans d'autres langues romanes (espagnol : Boyé & Cabredo, 2010 ; italien : Montermini & Boyé, 2012 ; catalan : Guerrero, 2014). Toutefois, avec l'apparition des bases de données lexicales et des traitements automatiques, ce type de modèle basé sur une notion de régularité catégorique déterminée de façon impressioniste a laissé la place à des analyses quantitatives cherchant à répondre à la question du remplissage des paradigmes (Ackerman et al., 2009) et à démontrer l'hypothèse de l'entropie faible des systèmes flexionnels.

Après avoir rappelé les grandes lignes des modélisations basées sur les espaces thématiques et des problèmes soulevés à leurs égards dans la section 2, nous proposons une transition vers un ensemble de données écologiques à la section 3. La section 4 présente les analogies qui serviront de base à notre nouveau modèle. La section 5 introduit la notion de prédictivité jointe qui sera utilisée dans la section 6 pour décrire le mécanisme utilisé pour remplir les paradigmes. Enfin, la section 7 décrit quelques modules supplémentaires pour définir ce nouveau modèle d'analyse flexionnelle quantitative basé sur des données écologiques dans la lignée des espaces thématiques.

1 Introduction

Pendant une dizaine d'années, la morphologie thématique a connu un succès pour les analyses flexionnelles et dérivationnelles en français (Bonami & Boyé, 2003, 2005, 2007, 2014 ; Bonami, Boyé, Giraudo & Voga, 2008 ; Bonami, Boyé & Kerleroux, 2009 ; Plénat 2008 ; Roché, 2010 ; Roché & Plénat, 2014 ; Tribout, 2010) et dans d'autres langues romanes (espagnol : Boyé & Cabredo, 2010 ; italien : Montermini & Boyé, 2012 ; catalan : Guerrero, 2014). Toutefois, avec l'apparition des bases de données lexicales et des traitements automatiques, ce type de modèle basé sur une notion de régularité catégorique, déterminée de façon impressionniste, a laissé la place à des analyses quantitatives cherchant à répondre à la question du remplissage des paradigmes (Ackerman *et al*., 2009) et à démontrer l'hypothèse de l'entropie faible des systèmes flexionnels[1].

Après avoir rappelé les grandes lignes des modélisations basées sur les espaces thématiques et des problèmes soulevés à leurs égards dans la section 2, nous proposons une transition vers un ensemble de données écologiques[2] à la section 3. La section 4 présente les analogies qui serviront de base à notre nouveau modèle. La section 5 introduit la notion de prédictivité jointe qui sera utilisée dans la section 6 pour décrire le mécanisme utilisé pour remplir les paradigmes. Enfin, la section 7 décrit quelques modules supplémentaires pour définir ce nouveau modèle d'analyse flexionnelle quantitative basé sur des données écologiques dans la lignée des espaces thématiques.

2 Les espaces thématiques

Bonami (2014, chapitre 2) présente en détail l'histoire des espaces thématiques, leur place dans les descriptions morphologiques du français dans les années 2000 et les problèmes rencontrés par ce type d'analyse. Nous reprenons ici quelques-uns des éléments principaux.

Au départ, les espaces thématiques sont un moyen de formaliser l'allomorphie radicale dans les paradigmes flexionnels. Il s'agit, d'une part, de capter la distribution des différents allomorphes dans les paradigmes (relations case-thème) et, d'autre part, de décrire les analogies entre allomorphes pour les lexèmes réguliers et la distribution de ces allomorphes au sein du paradigme flexionnel (relations thème-thème).

Le tableau 1 présente une analyse thématique typique de la conjugaison du français (12 thèmes indexés) avec une construction des formes basée sur un thème/radical et une désinence[3]. La ligne des formes non-finies (Inf/P. Prés./P. Passé) comprend la forme de l'infinitif, le participe présent, et les 4 formes des participes passés (M.SG, M.PL, F.SG, F.PL). Dans le cas général, l'analyse se présente plutôt comme un système de blocs PFM (Stump, 2001) avec des règles de sélection de thème (voir p.ex. Bonami & Boyé, 2007).

Cette analyse (Tab. 1) en combinaison avec les radicaux correspondant à chaque thème permet de conjuguer un verbe quelconque. Par exemple, le tableau 2, montre comment se réalise la conjugaison de BOIRE à partir de l'ensemble de ses thèmes.

Tableau 1. Flexion verbale basée sur l'espace thématique du français

	1SG	2SG	3SG	1PL	2PL	3PL
Présent	3	3	3	1+ɔ̃	1+e	2
Imparfait	1+ɛ	1+ɛ	1+ɛ	1+jɔ̃	1+je	1+ɛ
Futur	10+re	10+ra	10+ra	10+rɔ̃	10+re	10+rɔ̃
Conditionnel	10+rɛ	10+rɛ	10+rɛ	10+rjɔ̃	10+rje	10+rɛ
Subj. prés.	7	7	7	8+jɔ̃	8+je	7
Passé	H(11)	11	11	11+m	11+t	H(11)+r
Subj. imparf.	11+s	11+s	11	11+sjɔ̃	11+sje	11+s
Impératif	—	5	—	6+ɔ̃	6+e	—
Inf/P. prés./P. passé	9+r	4+ɑ̃	12-C	12-C	12	12

Tableau 2. Exemple de la conjugaison de BOIRE

	1SG	2SG	3SG	1PL	2PL	3PL
Présent	bwa	bwa	bwa	byv+ɔ̃	byv+e	bwav
Imparfait	byv+ɛ	byv+ɛ	byv+ɛ	byv+jɔ̃	byv+je	byv+ɛ
Futur	bwa+re	bwa+ra	bwa+ra	bwa+rɔ̃	bwa+re	bwa+rɔ̃
Conditionnel	bwa+rɛ	bwa+rɛ	bwa+rɛ	bwa+rjɔ̃	bwa+rje	bwa+rɛ
Subj. prés.	bwav	bwav	bwav	byv+jɔ̃	byv+je	bwav
Passé	H(by)	by	by	by+m	by+t	H(by)+r
Subj. imparf.	by+s	by+s	by	by+sjɔ̃	by+sje	by+s
Impératif	—	bwa	—	byv+ɔ̃	byv+e	—
Inf/P. prés./P. passé	bwa+r	byv+ɑ̃	by-C	by-C	by	by

1	byv	7	bwav
2	bwav	8	byv
3	bwa	9	bwa
4	byv	10	bwa
5	bwa	11	by
6	byv	12	by

En complément de ces relations case-thème, la description comprend un ensemble de relations qui permet de construire tous les thèmes des verbes « réguliers » à partir d'un thème donné.

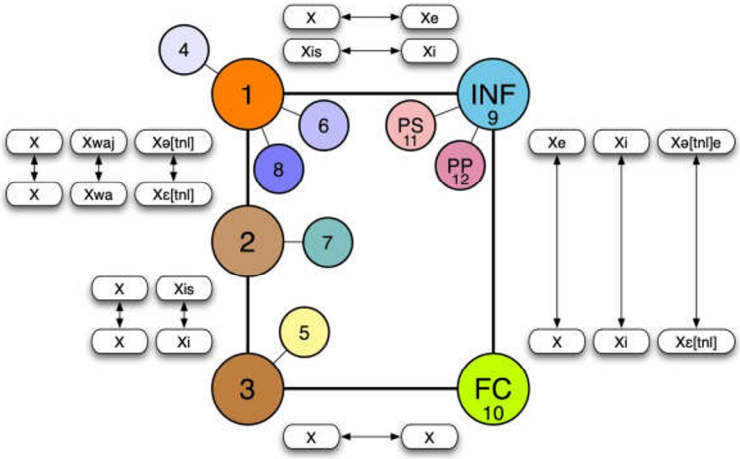


Figure 1. Graphe des relations régulières entre thèmes (Boyé, 2011)

La figure 1 montre les relations proposées par Boyé (2011). Ces relations captent les régularités des verbes irréguliers, autrement dit la configuration du graphe et les liens entre les thèmes captent la distribution des irrégularités dans le paradigme verbal. Les relations entre les thèmes représentent les transformations correspondant à différents types de verbes considérés réguliers pour l'analyse en question (laver, jeter, noyer, finir). Par exemple pour la correspondance entre thème 1 et thème 9, les 4 types de verbes considérés ont soit une transformation X-Xe (lav-lave, ʒət- ʒəte, nwaj-nwaje), soit une transformation Xis-Xi (finis-fini).

Pour BOIRE, la figure 2 montre comment les différents allomorphes sont distribués. Il y a trois sous-paradigmes qui correspondent respectivement à de la flexion régulière. La partie de la conjugaison de BOIRE qui correspond à °BUVER est représentée par tous les thèmes byv interconnectés en haut à gauche (1, 4, 6, 8). La partie qui correspond à °BOIVER par l'interconnexion entre 2 et 7 pour bwav et celle qui correspond à °BOYER, par les interconnexions des thèmes 3, 5 et 10 pour bwa.

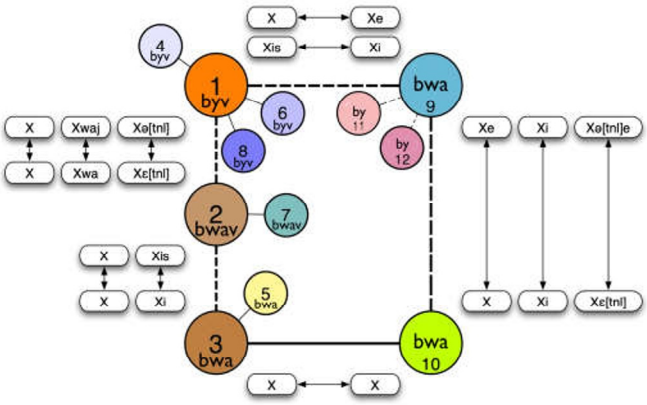


Figure 2. Graphe des relations pour BOIRE

Les thèmes 9 et 10 ne sont pas connectés car aucune des trois correspondances régulières (Xe-X, Xi-Xi, Xə[tnl]e- Xɛ[tnl]e) n'est respectée par bwa-bwa.

Bonami (2014) montre comment ce type d'analyse est suradaptée aux langues romanes où les classes flexionnelles traditionnellement convoquées pour décrire la conjugaison sont marginales par rapport à

celles rencontrées dans la déclinaison du latin ou du tchèque. De fait, le découpage présenté pour le français semble être basé sur un radical et des désinences mais une application de ce type d'analyse à une langue à classes flexionnelles canoniques ferait disparaître la plupart des « désinences » au profit des seules allomorphies « radicales », ce qui semble contre-intuitif. D'autre part, le choix des types de paradigmes flexionnels réguliers est arbitraire et ce choix joue un rôle fondamental dans la structuration du graphe des relations entre thèmes, aussi bien pour les places respectives des thèmes que pour les correspondances régulières entre eux.

Par exemple, sur la figure 2 comme expliqué précédemment, les thèmes 9 et 10 (infinitif et futur/conditionnel) ont la même valeur, ce qui est très souvent le cas pour les verbes irréguliers mais Boyé (2011) n'établit pas de relation entre eux pour ces cas d'homophonie. De la même façon, Boyé (2011) ne connecte pas les thèmes 11 et 12 malgré leurs homophonies fréquentes. Ces généralisations sont laissées de côté dans le cadre des espaces thématiques du fait d'une hypothèse fondamentale commune à toutes les analyses, la régularité est un phénomène qui concerne les classes flexionnelles et non des relations entre formes.

Malgré les problèmes soulevés par les analyses flexionnelles thématiques, elles permettent de fournir des moyens de préciser la notion de radical pour la morphologie constructionnelle. Avec l'analyse type présentée précédemment, Bonami & Boyé (2005) montrent que les adjectifs en -eur/-euse (p.ex. râleur/râleuse, buveur/buveuse) sont systématiquement construits sur le thème 1 du verbe correspondant (raler : 1=ral, boire :1=byv) et que les 2 radicaux des adjectifs correspondants (A et B) sont construits en parallèle l'un de l'autre (A=1+œr, B=1+øz).

Les analyses thématiques sont aussi une sorte de réponse précoce à la question du remplissage des paradigmes (Ackerman et al, 2009) puisqu'elles captent, pour les lexèmes « réguliers », le découpage d'une forme de départ et l'identification de son radical/thème au travers du tableau 1 dans notre exemple. À partir de ce radical, le graphe de la figure 1 permet d'obtenir tous les autres thèmes et de construire l'ensemble des formes comme dans le tableau 2.

3 Une transition écologique

Une grande partie des réponses initialement proposées à la question du remplissage des paradigmes, en commençant avec Ackerman *et al*. (2009), limitaient la portée du PCFP (Paradigm Cell Filling Problem) aux « nouveaux » lexèmes dans un cadre où toutes les formes des lexèmes « anciens » étaient connues. Ce type d'approche a été largement remis en cause par les études sur corpus (Blevins et al., 2017 ; Bonami & Beniamine, 2016) qui montrent que le nombre moyen de formes attestées par lexème reste faible. Pour un paradigme flexionnel comme celui des verbes en français, on ne trouve qu'environ 25 % des formes possibles attestées en corpus.

Pour analyser la question du remplissage des paradigmes de manière quantitative de façon réaliste (Boyé & Schalchli, 2019), nous avons choisi les données de Lexique4Linguist compilées par Schalchli (2022) à partir du corpus OpenSubtitle2016 (Lison & Tiedemann, 2016). Notre échantillon est donc basé sur des sous-titres en français de films avec l'idée que cette forme de texte proche de l'oral est une base raisonnable pour estimer les fréquences des formes verbales en milieu naturel.

Lexique4Linguist est un lexique qui contient environ 104.000 formes conjuguées pour 6.400 verbes. Les formes sont associées à leurs fréquences de token (comprises entre une et 6 millions d'occurences) et à une étiquette morphosyntaxique, les transcriptions phonétiques ont été copiées depuis Flexique (Bonami *et al*., 2013). Pour éliminer les bruits éventuels, nous avons limité notre corpus de travail aux formes ayant une fréquence supérieure à la médiane du corpus (occurences ≥ 6). Notre lexique de travail comprend 51.000 formes conjuguées et 4.800 verbes.

Comme le montre la figure 3[4], il y a une différence fondamentale entre notre lexique de travail et Flexique puisque ce dernier documente l'ensemble des formes fléchies pour tous les lexèmes tandis que, pour notre

lexique, il y a de grandes différences de représentation quantitative entre les cases du paradigme. Certaines sont pratiquement remplies pour tous les verbes, les cases de l'infinitif (inf), du participe passé masculin singulier (ppMS) et du présent indicatif 3sg (pi3S) sont remplies pour plus de 3500 verbes tandis que les cases de l'imparfait du subjonctif 1sg, 2sg, 1pl, 2pl, 3pl et du passé simple 2pl ne sont remplies que pour 10 verbes ou moins. Cette situation peut être mise en relation avec les difficultés des locuteurs avec certaines parties de la conjugaison.

Les différences sont aussi sensibles du point de vue des lexèmes. Comme le montre la figure 4, le nombre de cases remplies varie d'un lexème à l'autre : 600 verbes présentent une seule forme tandis que les 14 verbes ayant le plus de cases remplies comprennent entre 40 et 45 formes.

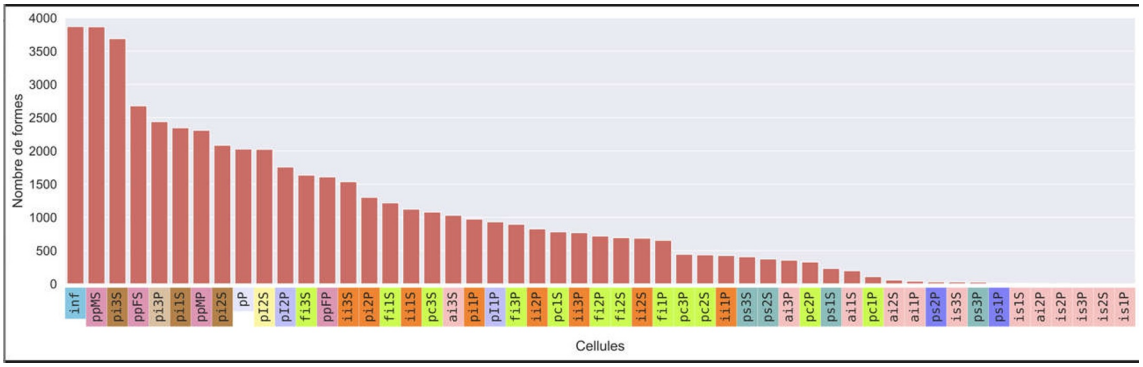


Figure 3. Nombre de formes par case du paradigme

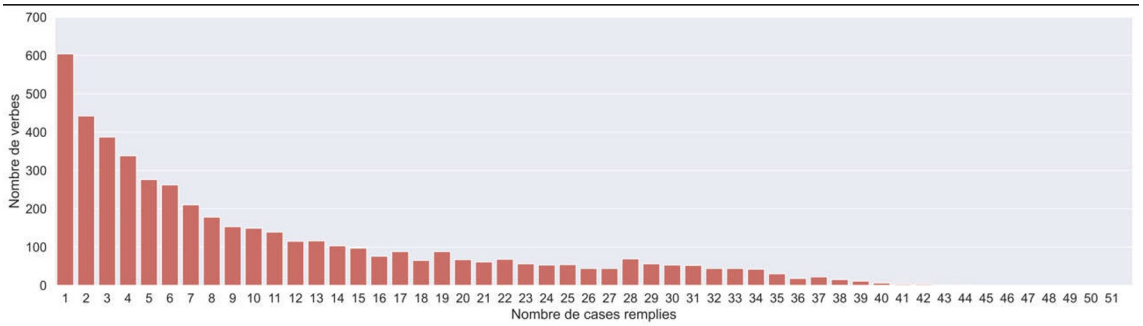


Figure 4. Nombre de verbes attestés par nombre de cases remplies

Sur la base de ce lexique de travail, on constate que la question du remplissage des paradigmes est loin de se limiter aux nouveaux lexèmes et qu'elle touche quasiment l'ensemble des lexèmes.

4 Des régularités aux analogies

À partir du lexique de travail, on va substituer aux relations par défaut entre thèmes, liées aux classes flexionnelles déclarées régulières dans le cadre des espaces thématiques, des analogies entre formes. Pour chaque paire de cases du paradigme, on calcule un ensemble d'analogies entre formes associées à partir de PredSPE (Bonami & Boyé, 2014).

Pour chaque paire de cases (a,b) du paradigme, PredSPE extrait toutes les paires des lexèmes attestés pour a et b et calcule les transformations qui font passer des formes connues de a aux formes connues de b. Ces analogies sont exprimées sous la forme de règles SPE (Chomsky & Halle, 1968) avec un contexte phonologique étroit[5].

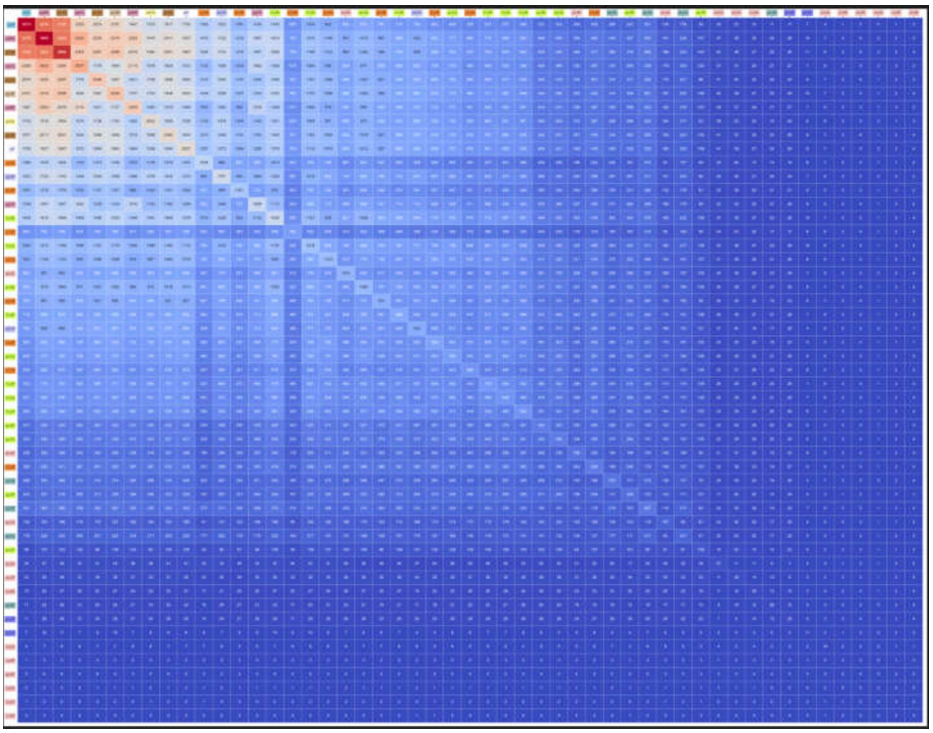
Le tableau 3 présente l'exemple des analogies entre imparfait 1PL et imparfait 1SG fourni par Bonami & Boyé (2014). Dans ce tableau, on trouve les trois transformations calculées pour cette paire, une transformation I sans contexte phonologique et deux transformations, II et III, chacune avec un contexte phonologique associé. La distribution des exemples entre les trois transformations est donnée dans la colonne probabilité : 84% pour I, 11% pour II, 5% pour III.

Tableau 3. Transformations entre imparfait 1PL et 1SG (Bonami & Boyé, 2014)

REGLE	EX. DE PAIRE	EX. DE LEXEME	EFFECTIF	PROBABILITE
I. jɔ̃→ɛ/__#	⟨kazjɔ̃,kazɛ⟩	CASER	5429	0,84
II. ɔ̃→ɛ/ [+cons, +haut, −cor] __#	⟨kadʁijɔ̃, kadʁijɛ⟩	QUADRILLER	722	0,11
III. ijɔ̃→ɛ/ [+cons, −voc, −son] [+cons, −voc, −nasal] __#	⟨kadʁijɔ̃, kadʁɛ⟩	CADRER	289	0,05

Les analogies abstraites de notre lexique sont plus limitées que celles de Bonami & Boyé (2014) qui reposaient sur un lexique fléchi complet. Leur nombre est plus petit et leurs contextes phonologiques sont moins étendus du fait de la plus petite quantité de formes associées à chaque case et par conséquent à chaque paire.

Tableau 4. Nombre de formes initiales remplies pour chaque paire de cases



Comme indiqué dans le tableau 4, suivant la paire considérée, le nombre de formes connues varie de 3.279 (orange foncé) pour la paire la plus représentée (infinitif, participe passé masculin singulier) à 10 ou moins (bleu foncé) pour toutes les paires contenant une forme de l'imparfait du subjonctif. Ces différences quantitatives produisent des différences qualitatives sur les analogies.

Le passage à un lexique extrait d'un corpus et à des analogies limitées à cet extrait constitue une transition écologique nécessaire pour une analyse réaliste, au sens de Boyé & Schalchli (2019), de la flexion verbale du français.

5 De la prédictivité simple à la prédictivité jointe

À partir des analogies obtenues, on applique la méthode de « classification des formes d'entrées » proposée par Bonami & Boyé (2014) pour regrouper les formes connues en fonction des analogies qui peuvent s'y appliquer. L'ensemble des formes convenant aux conditions phonologiques d'application d'un même ensemble d'analogies constitue une classe transformationnelle où chaque analogie de la classe est associée à un pourcentage d'utilisation au sein de la classe en question.

En continuant avec l'exemple précédent de Bonami & Boyé (2014) commencé dans le tableau 3, la classification des formes d'entrées donne les trois classes transformationnelles du tableau 5. A pour les formes convenant aux contextes des patrons I et II, B pour celles convenant à I, II et III, C pour celles ne convenant que à II.

Tableau 5. Classes transformationnelles entre imparfait 1PL et imparfait 1SG

CLASSE	PATRONS	EX. DE FORME	EX. DE LEXEME	EFFECTIF	PROBABILITE
A	{I,II}	kazjɔ̃, ʁasazjɔ̃	CASER, RASSASIER	5893	0,91
B	{I,II,III}	kadʁijɔ̃	QUADRILLER, CADRER	319	0,05
C	{II}	elwaɲɔ̃	ELOIGNER	228	0,04

Pour chacune de ces trois classes, les patrons sont associés à leur pourcentage d'utilisation pour les formes connues. Le tableau 6 donne cette répartition pour notre exemple. Pour la classe A avec les patrons {I, II}, la transformation I vaut pour 92% des cas et la II pour 8%. Pour la classe B avec les patrons {I, II, III}, la transformation I n'est jamais utilisée, la II vaut pour 9% des cas et la III pour 91%. Pour la classe C qui n'a qu'un seul patron II, c'est ce patron qui vaut dans 100% des cas.

Tableau 6. Probabilité conditionnelle d'utilisation des règles en fonction des classes

P(colonne|ligne)	I	II	III
A	0,92	0,08	0
B	0	0,09	0,91
C	0	1	0

Ces classes transformationnelles permettent de calculer la prédictivité d'une forme connue d'un lexème dans une case a pour le contenu d'une case b mais pas de calculer l'effet de la connaissance de plusieurs formes de départ pour ce même lexème.

Pour tenir compte des multiples formes connues d'un lexème dans le calcul la complexité du remplissage de son paradigme, Bonami & Beniamine (2016) propose une méthode basée sur ces classes transformationnelles : la prédictivité jointe (joint predictiveness). Pour la prédiction d'une forme f_S de la case S basée sur un ensemble des formes connues, par exemple les formes f_A, f_B, f_C des cases A, B, C, on combine les distributions associées aux transformations au niveau de chaque paire de {A, B, C, S}. En

utilisant ce calcul, on peut passer de la prédictivité des formes de départ à une prédiction des formes d'arrivée.

6 Du vote majoritaire au choix consensuel

Dans cette section, nous comparons deux méthodes pour remplir les paradigmes de notre échantillon à partir des classes transformationnelles calculées précédemment. Dans les deux cas, il s'agit de tenir compte des formes connues de l'échantillon et des classes transformationnelles calculées à partir de ces formes pour inférer les formes manquantes. La première stratégie consiste à remplir les paradigmes, lexème par lexème et case par case. Elle utilise directement les classes transformationnelles pour remplir le paradigme d'un lexème en calculant le meilleur candidat dans chaque case à partir des analogies sur l'ensemble des formes connues. La seconde procède également lexème par lexème mais cherche à produire un remplissage au niveau du paradigme complet en utilisant l'ensemble des analogies du système flexionnel.

6.1 Le vote majoritaire

En utilisant la prédictivité jointe, on peut prédire les formes manquantes d'un lexème en choisissant la forme la plus probable en fonction des formes connues. Ce mécanisme permet de remplir les paradigmes de chaque lexème attesté forme par forme. Dans l'ensemble, les résultats sont bons avec une précision de 97% et un rappel de 82% malgré les connaissances restreintes dont le système dispose avec 51.000 formes connues pour inférer les 170.000 formes manquantes des 4.800 verbes.

Toutefois, les erreurs présentes correspondent essentiellement à des remplissages incohérents. Le tableau 7 illustre le remplissage du paradigme de ABASOURDIR à partir de ses trois formes connues (participes passés, masculin singulier, masculin pluriel et féminin singulier). Dans le tableau, les formes connues sont notées en gras, les formes inférées correctes en noir et les erreurs en gris italique.

Tableau 7. Remplissage par vote majoritaire pour ABASOURDIR

Présent	abazuʁdi	abazuʁdi	abazuʁdi	*abazuʁdizɔ̃*	abazuʁdise	abazuʁdis
Imparfait	abazuʁdisɛ	abazuʁdisɛ	abazuʁdisɛ	abazuʁdisjɔ̃	*abazuʁdizje*	*abazuʁdizɛ*
Futur	abazuʁdiʁɛ	abazuʁdiʁa	abazuʁdiʁa	abazuʁdiʁɔ̃	abazuʁdiʁe	abazuʁdiʁɔ̃
Conditionnel	abazuʁdiʁɛ	abazuʁdiʁɛ	abazuʁdiʁɛ	abazuʁdiʁjɔ̃	abazuʁdiʁje	abazuʁdiʁɛ
Subj. prés.	abazuʁdis	*abazuʁdiz*	abazuʁdis	?	?	?
Passé	abazuʁdi	abazuʁdi	abazuʁdi	abazuʁdim	abazuʁdit	abazuʁdiʁ
Subj. imparf.	abazuʁdis	?	abazuʁdi	?	?	?
Impératif	—	abazuʁdi	—	*abazuʁdizɔ̃*	abazuʁdise	—
Inf/P. prés./P. passé	abazuʁdiʁ	abazuʁdisɔ̃	**abazuʁdi**	**abazuʁdi**	**abazuʁdi**	abazuʁdi

Les cinq erreurs sont cohérentes entre elles mais sont complètement déconnectées des autres formes inférées. Comme démontré par Bonami & Beniamine (2016), le fait d'utiliser plusieurs formes connues pour prédire une forme manquante permet de réduire l'entropie de façon significative dans de nombreux cas. Mais le choix majoritaire dérivé de ce type de calcul amène un certain nombre d'incohérences dans les paradigmes, sous la forme de classes flexionnelles inattestées, du fait que les inférences pour les différentes formes manquantes sont considérées comme indépendantes. Pour pallier ce problème, nous proposons un mouvement symétrique au passage de la prédictivité simple à la prédictivité jointe, et de prédire l'ensemble des formes conjointement en rendant les inférences dépendantes les unes des autres avec un choix consensuel.

6.2 Choix consensuel

Pour obtenir des ensembles d'inférences plausibles, la méthode du choix consensuel utilise une stratégie en trois temps pour identifier non plus des formes inférées indépendantes (f_1, f_2, …, f_n) qui sont cohérentes avec l'ensemble des formes connues mais un ensemble de formes inférées interdépendantes qui forme un tout cohérent avec les formes connues. Autrement dit, le remplissage du paradigme consiste à trouver un ensemble de formes cohérent avec les formes connues, c'est-à-dire toutes interconnectées entre elles. En théorie des graphes, un ensemble de nœuds où tous les membres sont connectés entre eux s'appelle une clique. Le but du choix consensuel est donc de trouver un groupe de formes connues et inférées qui forment une clique en termes de relation de transformation.

Dans la première phase, on applique les classes distributionnelles à l'ensemble du paradigme, comme pour le choix majoritaire mais en ne gardant, dans ces premières inférences, que les formes qui correspondent à toutes les formes connues. Puis, on réapplique les classes distributionnelles aux inférences de premier niveau pour trouver les ensembles de formes qui ont toutes des analogies entre elles, les cliques. En général, il y a un grand nombre de cliques pour chaque lexème. Pour répondre à la question du remplissage, il reste à trouver la meilleure clique. Dans le cas général, on propose d'utiliser les critères suivants par ordre d'importance décroissante :

- le remplissage doit être fidèle aux formes connues
- le remplissage doit être le plus complet possible (le plus de formes remplies)
- le remplissage doit être le plus cohérent possible (le meilleur score[6])

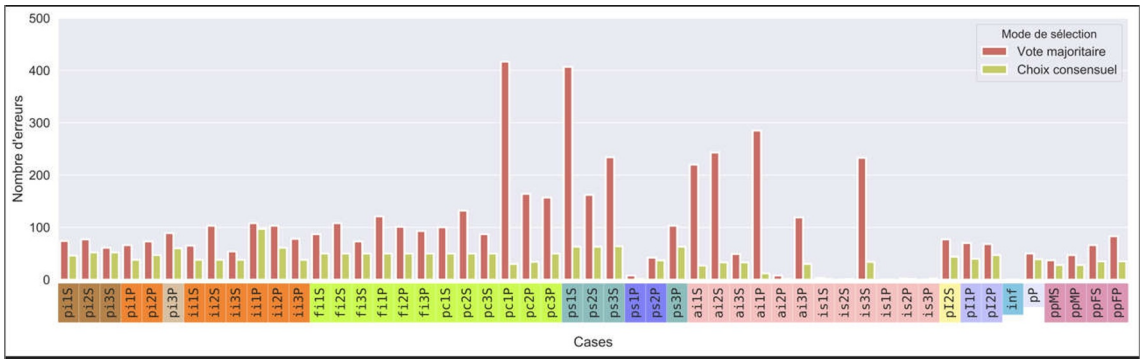


Figure 5. Nombre d'erreurs par case pour le vote majoritaire et le choix consensuel

Cette méthode permet d'obtenir des remplissages cohérents, correspondant à des classes flexionnelles plausibles contrairement au vote majoritaire. De fait, les résultats globaux sont sensiblement meilleurs avec une précision à 99% et un rappel à 82,3% liés à une forte baisse sur les pics d'erreurs du choix majoritaire comme le montre la figure 5.

On observe que, pour le remplissage de ABASOURDIR avec le choix consensuel illustré dans le tableau 8, les erreurs incohérentes du choix majoritaire du tableau 7 disparaissent.

Tableau 8. Résultat du choix consensuel pour ABASOURDIR

Présent	abazuʁdi	abazuʁdi	abazuʁdi	abazuʁdisɔ̃	abazuʁdise	abazuʁdis
Imparfait	abazuʁdisɛ	abazuʁdisɛ	abazuʁdisɛ	abazuʁdisjɔ̃	abazuʁdisje	abazuʁdisɛ
Futur	abazuʁdiʁɛ	abazuʁdiʁa	abazuʁdiʁa	abazuʁdiʁɔ̃	abazuʁdiʁe	abazuʁdiʁɔ̃
Conditionnel	abazuʁdiʁɛ	abazuʁdiʁɛ	abazuʁdiʁɛ	abazuʁdiʁjɔ̃	abazuʁdiʁje	abazuʁdiʁɛ
Subj. prés.	abazuʁdis	abazuʁdis	abazuʁdis	?	?	abazuʁdis
Passé	abazuʁdi	abazuʁdi	abazuʁdi	?	?	abazuʁdiʁ
Subj. imparf.	?	?	abazuʁdi	?	?	?
Impératif	—	abazuʁdi	—	abazuʁdisɔ̃	abazuʁdise	—
Inf/P. prés./P. passé	abazuʁdiʁ	abazuʁdisɑ̃	**abazuʁdi**	**abazuʁdi**	**abazuʁdi**	abazuʁdi

Il reste des problèmes de rappel pour le subjonctif présent 1PL, 2PL, le passé simple et l'imparfait du subjonctif. Ces problèmes sont en relation directe avec la sous-représentation de ces formes dans notre lexique de travail comme on peut le voir dans le tableau 9 qui reprend les données sur le nombre de verbes avec des formes attestées pour chacune des cases du paradigme de la figure 3.

Tableau 9. Nombre de formes par case du paradigme

Présent	2347	2084	3688	974	1301	2439
Imparfait	1123	686	1536	426	826	769
Futur	1218	694	1635	654	719	896
Conditionnel	782	436	1080	108	329	444
Subj. prés.	231	373	407	*11*	*23*	*20*
Passé	197	57	1031	*33*	*0*	356
Subj. imparf.	*10*	*3*	*22*	*2*	*4*	*3*
Impératif	—	2023	—	932	1757	—
Inf/P. prés./P. passé	3870	2027	3866	2309	2677	1608

Alors que la sous-représentation des formes du passé simple et de l'imparfait du subjonctif semblent naturelles, celle du subjonctif présent au pluriel semble plus tenir d'une erreur systématique au moment de l'étiquetage automatique des sous-titres. Pour corriger ce problème, il faudrait sans doute revoir ce processus. Toutefois, en dépit de la très faible représentation du subjonctif présent 3PL, le fait qu'il appartienne à la même zone thématique que le subjonctif présent singulier permet au modèle de rappeler la forme correctement comme on va le voir dans la section suivante.

7 Le petit monde de la flexion[7]

Bien que le principe du choix consensuel appliqué à un ensemble écologique de formes soit l'élément fondamental du modèle présenté ici, il ne constitue qu'une première partie du processus qui repose également sur trois autres composants.

7.1 Le paradigme morphomique

À partir de la classification morphosyntaxique fournie par l'étiquetage automatique, on fait émerger un paradigme morphomique[8] en fusionnant des ensembles de cases. Une paire de cases est fusionnable si toutes les formes attestées pour les deux cases sont identiques avec au moins un cas où les deux formes sont attestées pour le même lexème. Pour fusionner un ensemble comme les quatre cases du subjonctif présent 1SG, 2SG, 3SG, 3PL, il ne suffit pas que la première soit fusionnable avec la seconde, la seconde avec la

troisième et la troisième avec la quatrième. Il faut qu'elles soient toutes fusionnables entre elles, qu'elles forment une clique pour la relation de fusionnabilité.

Dans le cas de notre lexique de travail, on obtient les regroupements suivants : pi2S/pi3P, ii1S/ii2S/ii3S/ii3P, fi1S/pc1S/pc2S/pc3S/pc3P, fi2S/fi3S, fi1P/fi3P, ps1S/ps2S/ps3S/ps3P, ai2S/ai3S/is3S, is1S/is2S/is3P, ppMS/ppMP, ppFS/ppFP

Le passage à un paradigme morphomique permet de repeupler immédiatement les cases puisque la première prédiction consiste à partager les formes entre les cases membres des regroupements. Ce qui donne la différence entre le remplissage syntaxique initial et le remplissage morphomique dans la figure 6 permettant ainsi des prédictions efficaces pour le subjonctif présent 3PL en dépit du petit nombre de formes représentées initialement.

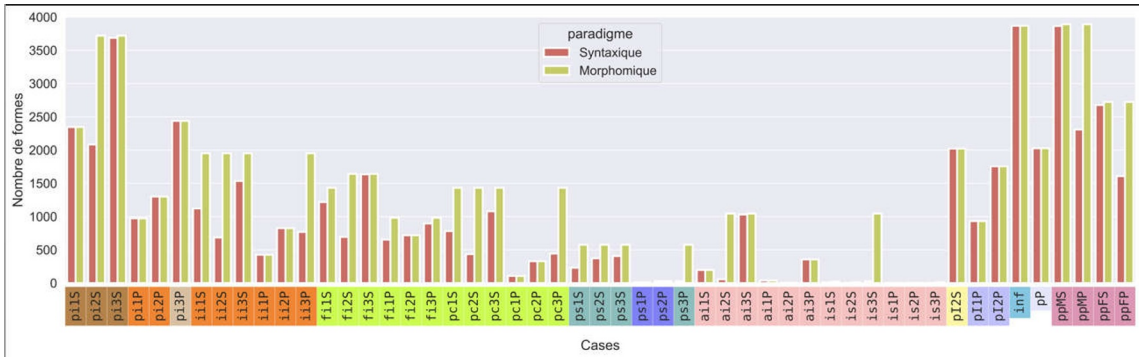


Figure 6. Remplissage initial en fonction du type de paradigme

7.2 Le choix consensuel à deux tours

Malgré la bonne précision (99,2%) des résultats initiaux du choix consensuel à 1 tour, leur rappel est relativement faible à cette étape du processus (76,8%). Pour améliorer celui-ci, on procède à un deuxième tour basé sur les données obtenues au premier tour et les mêmes transformations mais en supprimant les contraintes phonologiques[9]. Les formes inférées du premier tour imposant plus de contraintes de cohérence sur les cliques de formes à inférer au deuxième tour, celles-ci permettent de compenser l'abandon des contraintes phonologiques et de maintenir un niveau de précision proche du premier tour (99%) tout en gagnant en rappel (81.2%).

7.3 Le retour des espaces thématiques

Du fait de l'usage systématique des cliques pour choisir les formes inférées, il se peut que les formes manquantes après le deuxième tour de choix consensuel se trouvent à l'intérieur d'une zone d'interprédiction des espaces thématiques traditionnels. En effet, la contrainte de sélection des formes impose qu'il y ait des relations entre toutes les formes inférées et connues mais, dans un contexte écologique, il est possible que certaines relations ne soient pas représentées dans les données. Cependant, il est possible de prédire encore une partie des formes manquantes en faisant appel aux transformations à entropie nulle comme celles associées à la relation du futur 3SG vers le futur 3PL dotée d'une seule transformation Xa => Xɔ̃.

Parmi ces transformations, il y a des différences de fiabilité. Certaines ont une entropie nulle parce qu'elles sont basées sur tellement peu d'exemples que les différentes alternatives ne sont pas représentées. Pour éviter ces transformations, on fixe un seuil sur le nombre de formes attestées pour la paire concernée (p.ex 10% du nombre de lexèmes). Sur cette base, on peut remplir les formes manquantes et augmenter encore le rappel (82,3%) sans changer la précision (99%).

8 Conclusion

20 ans après, le type de modèle proposé ici tente d'éviter les écueils majeurs des espaces thématiques :

- une base de connaissance encyclopédique
- une définition arbitraire de la régularité
- un graphe de relation trop restreint

La transition écologique fait que les connaissances comme les inférences sont abstraites depuis un corpus. L'association des formes connues avec leurs propriétés syntaxiques reste un point opaque pour lequel on peut chercher des solutions avec la sémantique distributionnelle et des comparaisons de vecteurs pour différentes formes fléchies[10].

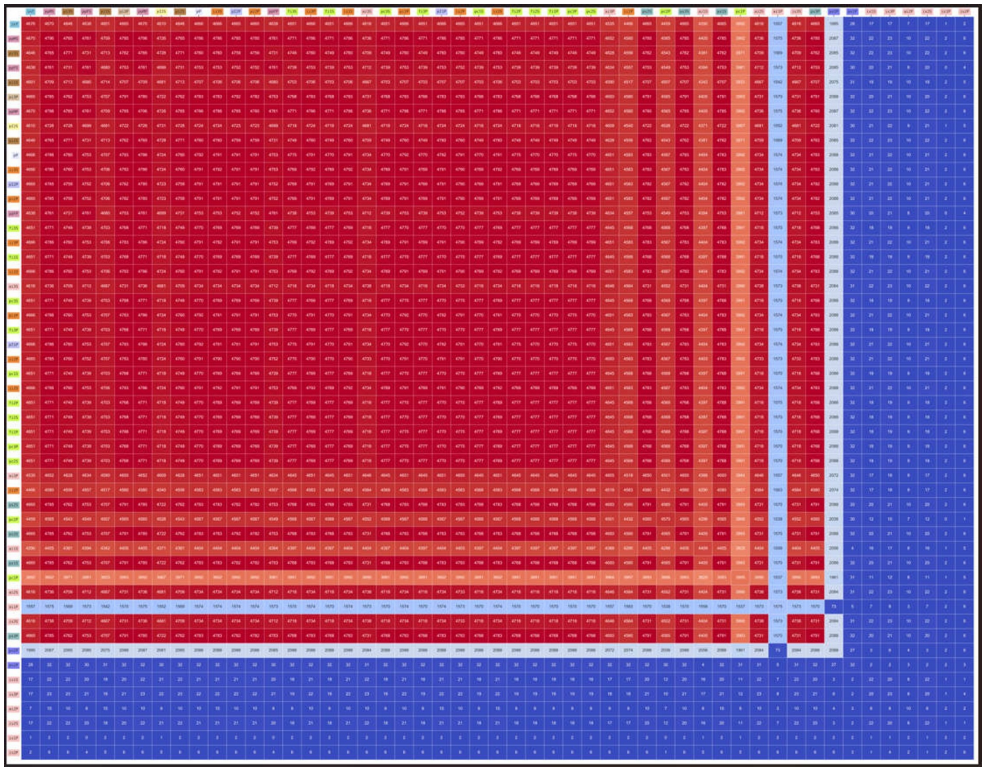
La notion de régularité au niveau des paradigmes complets cède la place à des analogies locales entre formes. Les contextes d'application des analogies sont directement étendus en fonction de leurs champs d'application. Les analogies ne concernant qu'un lexème unique sont comptabilisées mais pas généralisées.

Avec un graphe où toutes les formes sont interconnectées, les patrons flexionnels émergent du calcul des cliques. Les cliques qui couvrent entièrement le paradigme flexionnel correspondent aux classes flexionnelles traditionnelles. La régularité émerge directement du processus de remplissage des paradigmes en fournissant des résultats relativement complets pour les « nouveaux » lexèmes comme pour les « anciens ».

Dans cette optique écologique, on peut comparer le tableau de l'état initial des formes remplies pour les paires de cases dans le lexique de travail (Tab. 4) avec le résultat final (Tab. 10) qui représenterait pour la première l'exposition réelle d'un locuteur et pour la deuxième les connaissances perçues par le locuteur.

Il reste une zone défective qui correspond aux cases peu représentées dans le lexique initial et qui sont la cause des problèmes de rappel. Pour la partie hors passé simple et imparfait du subjonctif, la précision est quasiment la même (99%) mais le rappel est nettement supérieur (94,7% contre 82,3%). La défectivité a une place particulière puisque tous les lexèmes sont « défectifs » mais leurs formes manquantes seraient inférées par les locuteurs et par le modèle. Pour les lexèmes défectifs au sens classique, le problème n'est pas seulement que leurs formes manquantes ne sont pas attestées puisque beaucoup d'autres lexèmes sont dans ce cas, mais plutôt que leurs formes manquantes ne sont pas acceptées[11].

Tableau 10. Nombre de formes finales remplies pour chaque paire de cases



Pour l'interface avec la morphologie constructionnelle, la nouvelle configuration ne change pas fondamentalement les descriptions. Les références utilisées dans les analyses thématiques sont remplaçables par des références directes à des formes. Pour la construction des adjectifs déverbaux de (Bonami & Boyé, 2005), par exemple, la dérivation basée sur le thème 1 peut être exprimée en fonction de l'imparfait 3SG, et au lieu de construire un thème A et un thème B pour l'adjectif, il suffirait de construire une forme M.SG et une forme F.SG.

D'une façon plus générale, la notion de clique joue un rôle central dans ce modèle, tant pour constituer les paradigmes que pour leur remplissage. Dans le premier cas, elle permet de faire émerger le paradigme morphomique et, dans le deuxième, elle constitue les ensembles de formes consensuelles. Les paradigmes complets ainsi constitués reproduisent les classes flexionnelles arbitrairement constituées au travers des connaissances encyclopédiques.

Pour terminer, d'autres paramètres pourraient être pris en compte pour le calcul des distributions au sein des classes de transformation comme le genre dans les langues slaves qui joue un rôle crucial dans les choix des cliques flexionnelles pour les noms, où la transitivité pour les langues avec accord sujet et objet dont le paradigme est à géométrie variable.

Références bibliographiques

Ackerman, F., Blevins, J. P., & Malouf, R. (2009). Parts and wholes: implicative patterns in inflectional paradigms. *Analogy in Grammar.* pages 54–82. Oxford : Oxford University Press.

Albright, A. (2002). *The identification of bases in morphological paradigms*. Ph.D. thesis, UCLA.

Blevins, J. P., Milin, P., & Ramscar, M. (2017). The zipfian paradigm cell filling problem. *Perspectives on morphological organization*, pages 139–158. Brill.

Bonami, O. (2014). *La structure fine des paradigmes de flexion*. HDR, Université Paris 7 - Denis Diderot.

Bonami, O. & Beniamine, S. (2016). Joint predictiveness in inflectional paradigms. *Word Structure, 9(2)*, 156–182.

Bonami, O. & Boyé, G. (2003). Supplétion et classes flexionnelles dans la conjugaison du français. *Langages, 152*, 102–126.

Bonami, O. & Boyé, G. (2005). Construire le paradigme d'un adjectif. *Recherches linguistiques de Vincennes, 34*, 77–98.

Bonami, O. & Boyé, G. (2007). French pronominal clitics and the design of paradigm function morphology. *On-line Proceedings fo the Fith Mediterranean Morphology Meeting (MMM5)* Fréjus 15-18 September 2005, pages 291–322. Bologna : Università degli Studi di Bologna.

Bonami, O. & Boyé, G. (2014). De formes en thèmes. *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, pages 17–45. Nanterre : Presses Universitaires de l'Université Paris Nanterre.

Bonami, O., Boyé, G., Giraudo, H., & Voga, M. (2008). Quels verbes sont réguliers en français? *Actes du premier Congrès Mondial de Linguistique Française*, pages 1499–1511. Paris : ILF / EDP Science.

Bonami, O., Boyé, G., & Kerleroux, F. (2009). L'allomorphie radicale et la relation flexion-construction. *Aperçus de morphologie du français*, pages 267–286. Saint-Denis : Presses Universitaires de Vincennes.

Bonami, O., Caron, G., & Plancq, C. (2013). Flexique: an inflectional lexicon for spoken french. Technical documentation [http://www.llf.cnrs.fr/flexique/documentation. pdf].

Boyé, G. (2011). Régularités et classes flexionnelles dans la conjugaison du français. *Des unités morphologiques au lexique*, pages 41–68. Hermes Science Publishing.

Boyé, G. & Cabredo Hofherr, P. (2010). Defectiveness as stem suppletion in French and Spanish verbs. *Defective Paradigms: Missing forms and what they tell us*, pages 35–52, Oxford : The British Academy, Oxford University Press.

Boyé, G. & Schalchli, G. (2016). The status of paradigms. *The Cambridge Handbook of Morphology*, pages 206–234, Cambridge: Cambridge University Press.

Boyé, G. & Schalchli, G. (2019). Realistic data and paradigms: the paradigm cell finding problem. *Morphology, 29(2),* 199–248.

Chomsky, N. & Halle, M. (1968). *The Sound Patterns of English*. Cambridge: M.I.T. Press.

Guerrero, A. (2014). Analyse thématique de la flexion en catalan central standard. Thèse de doctorat, Université Toulouse le Mirail - Toulouse II.

Lison, P. & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781, 2013

Montermini, F. & Boyé, G. (2012). Stem relations and inflection class assignment in Italian. *Word Structure, 5*, 69–87.

Plénat, M. (2008). Le thème L de l'adjectif et du nom. *Actes du premier Congrès Mondial de Linguistique Française*, pages 1613–1626. Paris : ILF / EDP Science.

Rajman, M., Lecomte, J., & Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Technical Report GRACE, GTR 3-2.1, LIMSI.

Roché, M. (2010). Base, thème, radical. *Recherches Linguistiques de Vincennes, 39*, 95–134.

Roché, M. & Plénat, M. (2014). Le jeu des contraintes dans la sélection du thème présuffixal. *Actes du quatrième Congrès Mondial de Linguistique Française*. Paris : ILF / EDP Science.

Schalchli, G. (2022). Lexique4linguists. ORTOLANG (Open Resources and TOols for LANGuage).

Shannon., C. E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423 & 623–656.

Stump, G. (2001). *Inflectional Morphology. A Theory of Paradigm Structure*. Cambridge: Cambridge University Press

Tribout, D. (2010). *Les conversions de nom à verbe et de verbe à nom en français*. Thèse de doctorat, Université Paris Diderot.

---

[1] En théorie de l'information (Shannon, 1948), l'entropie mesure l'incertitude sur l'information dans un système. Dans les systèmes flexionnels, l'entropie est utilisée pour quantifier les probabilités d'erreur sur la prédiction des formes fléchies. Une entropie nulle correspond à une prédiction certaine, une entropie de 1 à une prédiction fiable à 50 %. L'hypothèse de l'entropie faible des systèmes flexionnels signifie que les systèmes flexionnels seraient très prédictibles.

[2] Les données écologiques sont celles qui seraient vraisemblablement accessibles aux locuteurs. Elles sont en général basées sur des corpus et associées à des données de fréquences.

[3] Pour faciliter le répérage des thèmes, ils sont colorés dans ce tableau et les suivants.

[4] Les étiquettes Grace (Rajman *et al*, 1997) utilisées pour les noms des cases flexionnelles sont basées pour les formes finies sur une première lettre pour le temps (p : présent, i : imparfait, a : passé, f : futur), une deuxième lettre pour le mode (i : indicatif, s: subjonctif, c : conditionnel, I : impératif), un numéro de personne (1, 2, 3) et un nombre (sg : S, pl : P). Pour les formes non-finies, on a inf pour infinitif, pP pour participe présent et pp suivi de MS, MP, FS, FP pour les différentes formes de participe passé.

[5] PredSPE comme le MGL de Albright (2002) étend les contextes phonologiques des analogies au minimum naturellement possible, la différence majeure étant que PredSPE ne génère qu'une seule règle pour chaque transformation.

[6] Les cliques reçoivent un score qui correspond à la force des votes entre les différentes formes de la clique.

[7] Le nom du modèle proposé est SWIM : Small World Inflectional Morphology.

[8] Un paradigme morphomique (Boyé & Schalchli, 2016) diffère d'un paradigme morphosyntaxique en regroupant les cases qui sont systématiquement syncrétiques pour tous les lexèmes.

[9] Un grand merci à Berthold Crysmann pour cette suggestion.

[10] Les vecteurs de mots (Mikolov et al., 2013) permettent de situer les mots d'un corpus dans un espace sémantique organisé où les mots sont d'autant plus proches qu'ils apparaissent dans les mêmes contextes.

[11] Pour une discussion des problèmes d'acceptation des formes défectives, voir Copot & Sims (à paraître)