SHS Web of Conferences **27**, 08001 (2016) DOI: 10.1051/shsconf/20162708001
Congrès Mondial de Linguistique Française - CMLF 2016

Pour une modélisation surfaciste de la flexion
Le cas de la conjugaison du français

Gilles Boyé

Université Bordeaux-Montaigne (UBM) (France)
Cognition, Langues, Langage, Ergonomie (CLLE-ERSSàB) - UMR5263 CNRS - Université Michel de Montaigne - Bordeaux III (France)

gilles.boye@u-bordeaux-montaigne.fr

Résumé. Les descriptions de la morphologie flexionnelle des langues sont naturellement basées sur l'observation détaillée de l'ensemble des propriétés connues des systèmes en question mais quelle part de ces connaissances est accessible aux locuteurs ? Dans ce travail nous proposons une approche résolument basée sur des connaissances accessibles illustrée sur la conjugaison du français. Le modèle proposé ici vise à répondre à la question du remplissage des paradigmes (Paradigm Cell Filling Problem : Ackerman et al., 2009) à partir des connaissances lexicales présentes dans un lexique-échantillon et de l'inférence de connaissances flexionnelles dans le cadre de la théorie de l'information (Shannon, 1948). Notre analyse est basée sur l'observation classique que un paradigme flexionnel est un ensemble de formes toutes mutuellement reliées par des analogies. À partir de ce point, nous proposons un mécanisme de construction des connaissances flexionnelles et de remplissage des paradigmes.

Absract. Inflectional morphology descriptions are usually based on fine-grained observation of the complete set of forms and known properties of the systems in question but what part of this knowledge could actually be accessible to speakers? In this study, we propose a data driven approach of inflectional morphology based on available knowledge applied to French conjugation. Our model aims at solving the Paradigm Cell Filling Problem (Ackerman & al., 2009) starting from partial inflectional knowledge, a sparsely populated organized lexicon and builds up inflectional paradigms through analogies and network connections using Information Theory (Shannon, 1948). Our analysis relies on the classical observation that forms in an inflectional paradigm are related by mutual analogies. Theses analogy sets constituting inflectional classes. What we propose here is a mechanism for the emergence of inflectional classes.

1 Pour une approche surfaciste

Stump (2001) définit quatre type d'approches pour la morphologie flexionnelle avec deux oppositions devenues classiques :

— les *approches lexicales* où les informations flexionnelles sont associées à des morphes vs les *approches inférentielles* où les informations flexionnelles sont associées à des transformations phonologiques

— les *approches incrémentales* qui subordonnent les traits de flexion à la présence des exposants vs les *approches réalisationnelles* qui subordonnent la présence des exposants aux traits de flexion

Au sein des approches inférentielles réalisationnelles comme A-Morphous Morphology (Anderson, 1992), Network Morphology (Corbett & Fraser, 1993; Brown & Hippisley, 2012), Paradigm Function Morphology (Stump, 2001) ou Natural Morphology (Kilani-Schoch & Dressler, 2005), Blevins (2006) distingue deux autres perspectives :

— une approche constructive examine l'ensemble du paradigme d'un lexème pour en déduire les éléments syntagmatiques de la flexion et construire les différentes formes fléchies à partir de ces éléments

— une approche abstractive relie les formes fléchies entre elles sans chercher à les réduire à des éléments linguistiques abstraits plus petits

La critique des approches constructives de Blevins se base sur le fait qu'elles utilisent des éléments abstraits potentiellement inaccessibles aux locuteurs natifs, les radicaux et les exposants, dont la postulation n'est pas nécessaire.

Nous proposons ici d'étendre cette critique à toutes les approches qui reposent sur une connaissance totale du lexique et du système flexionnel dans son ensemble. En effet, il est peu probable que les locuteurs du français, par exemple, disposent d'une représentation complète de toutes les conjugaisons de tous les verbes, ni même de toutes les classes flexionnelles [1] comme supposé par certains travaux récents sur ce sujet (Stump & Finkel, 2013; Bonami & Beniamine, 2015; Beniamine & Sagot, 2015).

1. L'existence des dictionnaires de verbes comme le Bescherelle (Arrivé, 1997) montre les difficultés des locuteurs natifs avec la conjugaison de leur langue.

Le modèle proposé ici vise à répondre à la question du remplissage des paradigmes (Paradigm Cell Filling Problem : Ackerman et al., 2009) à partir des connaissances lexicales présentes dans les lexiques-échantillons dans le cadre de la théorie de l'information (Shannon, 1948).

Notre analyse est basée sur une observation classique :

(1) un paradigme flexionnel est un ensemble de formes toutes mutuellement reliées par des analogies [2]

Le constat en (1) est celui qu'on trouve à la racine de la définition formelle des classes flexionnelles, par exemple dans (Beniamine & Sagot, 2015). C'est ce même principe qui guide les étapes du processus décrit dans la suite de cette section. La première étape se concentre sur la recherche des *analogies*, on note toutes celles existant entre les formes présentes dans les lexiques-échantillons et qui définissent les relations dans des paradigmes attestés. Dans la seconde étape, on remplit les paradigmes en ajoutant toutes les *formes* reliables par analogies aux formes connues créant une surabondance massive dont on extrait des ensembles de formes **toutes mutuellement** reliées par des analogies pour récolter les *paradigmes* flexionnels au sein des surabondances.

2 La modélisation des connaissances

Pour cette analyse basée sur la théorie de l'information, nous commençons par expliciter nos hypothèses sur les connaissances dont disposent les locuteurs : des connaissances lexicales et des connaissances sur la conjugaison.

2.1 Modélisation des connaissances lexicales

Les connaissances lexicales d'une vingtaine de pseudo-locuteurs différents [3] sont modélisées par des lexiques-échantillons constitués par tirage aléatoire de 5% dans un lexique de référence considéré l'ensemble total des connaissances lexicales possibles (15 000 formes sur environ 300 000).

2.1.1 Le lexique de référence

Notre lexique de référence est basé sur BDLex [4] (de Calmès & Pérennou, 1998) auquel ont été ajoutées des informations de fréquences de lexèmes et de formes de Lexique3 (New et al., 2001).

Les paradigmes flexionnels sont formés directement à partir des informations fournies par BDLex. Pour chaque lexème, les 51 cases sont peuplées par leurs formes étiquetées avec les propriétés flexionnelles correspondantes. Le regroupement aboutit à des paradigmes qui ne contiennent pas tous 51 formes pour les 51 cases. On observe 1 690 cases avec plus

2. Cette observation concerne de prime abord les paradigmes réguliers, nous l'étendons ici à l'ensemble des paradigmes flexionnels. Dans ce contexte, les analogies sont à comprendre comme des règles de transformation entre formes. Une analogie entre lavɔ̃ et lave permet de passer de la première forme à la deuxième en transformant le ɔ̃ final et e.

3. La discussion se limitera ici à la description de la modélisation elle-même. Nous supposons par ailleurs à la suite de Gaume et al. (2014) que des réseaux lexicaux différents donnent lieu aux mêmes métagénéralisations au travers du traitement proposé ici. Les résultats généraux fournis ici rendent compte des moyennes obtenues pour les 20 tirages aléatoires de lexiques-échantillons.

4. 6 561 verbes, 328 103 formes fléchies avec des révisions mineures de certaines transcriptions phonologiques pour améliorer la cohérence interne du lexique.

d'une forme pour des verbes surabondants (ASSEOIR, BALAYER, ABOYER, ...) et 6843 cases vides dont 2358 correspondent aux 3 formes des participes passés qui ne sont pas renseignées pour les 786 verbes intransitifs qui sélectionnent l'auxiliaire AVOIR (BADINER, ORBITER, VACILLER), 3000 pour 60 verbes vestiges qui ne possèdent qu'une forme dans BDLex (ACCROIRE, FÉRIR, CHAUVIR, ...) et les 1485 restantes pour des verbes défectifs (PLEUVOIR, FALLOIR, DISTRAIRE, ...).

Les fréquences utilisées sont issues de Lexique3, ce qui pose plusieurs problèmes :

(2) a. tous les lexèmes de BDLex ne sont pas présents dans Lexique3
b. de nombreuses formes possèdent une fréquence pour le lexème mais pas de fréquence propre pour la forme
c. Lexique3 donne une fréquence globale pour les formes homographes d'un même lexème [5]. Par exemple, il n'y a pas d'informations sur la distribution entre indicatif présent et subjonctif présent dans les cas de neutralisation graphique.

Pour obtenir des indices de fréquence sur l'ensemble des formes verbales de BDLex, nous avons calculé l'indice en (3) à partir des fréquences fournies par Lexique3 pour les deux bases de données (Frantext et Films) en combinant les fréquences des lexèmes (freqlemlivre, freqlemfilms2) et les fréquences de formes (freqlivre, freqfilms2).

(3) indice = 1 000 000x(freqlivre+freqfilms2)+(freqlemlivre+freqlemfilms2)+0,01

Cet indice permet de prendre en compte, la fréquence des formes et la fréquence des lexèmes quand elles sont disponibles. Pour les formes qui n'ont pas de fréquence propre, le facteur concernant le lexème permet de différencier ces dernières de celles des lexèmes qui n'apparaissent pas du tout dans Lexique3 et qui sont considérées comme encore beaucoup moins fréquentes [6]. Pour les formes neutralisées (2c), nous avons, pour l'instant, reporté la même fréquence pour toutes les formes homographes.

2.1.2 Les lexiques-échantillons

La constitution des lexiques-échantillons représentatifs de formes connues par les locuteurs est basée sur un tirage aléatoire de formes. La probabilité qu'une forme appartienne à un lexique-échantillon est proportionnelle à sa fréquence. Ce type d'échantillon est prévu pour représenter une part vraisemblable de connaissances lexicales.

Les 100 verbes les plus fréquents ont tous plus de 15 formes représentées dans l'échantillon et les plus fréquents possèdent des paradigmes presque complets (FAIRE 47, AVOIR 47, ÊTRE 45, DIRE 42, DEVOIR 42, ALLER 41, VOIR 41). Pour les verbes de fréquences faibles, 57% n'ont aucune forme qui ait été tirée au hasard dans l'échantillon, 13% de verbes n'ont reçu qu'une seule forme dans l'échantillon. Aucune généralisation pour les analogies ne peut être calculée à partir d'une seule forme mais ces verbes sont pris en compte pour la génération des paradigmes. La figure 1 montre la distribution au sein de nos lexiques-échantillons de 15 000 formes. On y voit clairement que la plupart des verbes représentés dans les lexiques-échantillons ne possèdent que très peu de formes chacun, plus de la moitié des verbes comptent moins de 4 formes.

Les premiers échantillons contiennent environ 1 150 verbes de basses fréquences avec entre 2 et 7 formes.

5. En revanche, les formes homographes de lexèmes de catégories différentes possèdent des fréquences distinctes.

6. Le facteur 0,01 représente l'idée que ces formes sont moins fréquentes que celles apparaissant dans Lexique3 d'au moins un ordre de grandeur, tout en permettant le tirage aléatoire de formes absentes de Lexique3 dans une proportion très minoritaire dans les échantillons.

quartiles	25%	50%	75%	90%	100%
nb formes	1	3	7	15	48
nb lexemes	863	1271	2015	2556	2800

TABLE 1 – Quartiles du nombre de formes par lexème

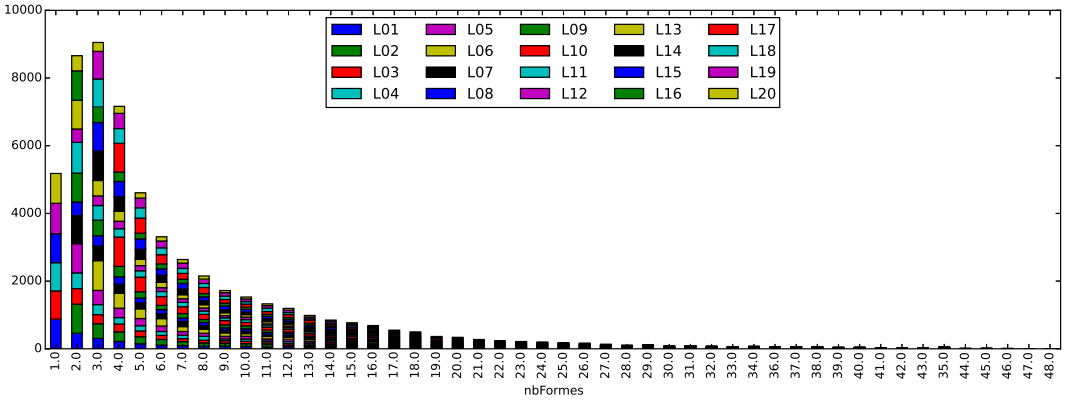


FIGURE 1 – Distribution du nombre de formes par lexème (20 échantillons)

2.2 Modélisation des connaissances flexionnelles

Au delà des connaissances lexicales, nous modélisons les connaissances flexionnelles en utilisant deux niveaux de description :

— des correspondances analogiques entre paires de formes fléchies conditionnées par leurs contextes phonologiques

— une répartition statistique entre les correspondances en compétition pour les mêmes contextes.

Nous utilisons la procédure introduite par Bonami & Boyé (2014) et reprise notamment par Bonami & Luís (2014) pour étudier les rapports entre chaque paire de cases du paradigme.

2.2.1 Analogies

Dans un premier temps, on calcule des règles de transformation analogiques (RTA) entre formes avec une version adaptée du Minimal Generalization Learner (MGL : Albright, 2002; Albright & Hayes, 2003) pour calculer le contexte minimalement général de chaque analogie. Contrairement au MGL original, nous ne conservons que la RTA la plus générale pour chaque analogie, il n'y a pas de mémoire masquée des informations lexicales [7].

La deuxième étape a pour but de remplacer la notion de confiance, utilisée par Albright & Hayes [8], liée à chaque règle isolée par une information sur la distribution des formes existantes dans les cas de compétitions entre RTA pour une même forme de départ. Pour chaque forme de départ, on note les RTA dont le contexte correspond à cette forme et on

7. Le MGL de Albright & Hayes conserve toutes les informations lexicales pour effectuer toutes les combinaisons possibles entre paires de formes et calculer toutes les généralisations intermédiaires possibles. Ici, les analogies sont généralisées systématiquement à chaque nouvelle paire sans conserver les informations précédentes.

8. La confiance associée à une RTA est le ratio entre le nombre de formes susceptibles d'être transformées par la règle et le nombre de formes correspondant effectivement à cette transformation.

constitue une classe de compétition. Pour une classe de compétition, on utilise les formes connues en sortie pour déterminer la répartition statistique des formes d'entrée entre les différentes RTA [9].

Pour le passage de la forme du présent indicatif 3sg à celle du présent indicatif 3pl (pr3sg ⟶ pr3pl), une des RTA consiste à ajouter un s à la finale de la forme d'entrée :

fini⟶finis

La généralisation minimale appliquée à notre ensemble de formes connues permet de limiter le contexte de cette RTA, ici, par exemple, aux formes se terminant par une voyelle d'avant non-arrondie :

Ø⟶s / X[iɛa] —

2.2.2 Classes de compétition

Les RTA de chaque paire de cases se trouvent en compétition pour certains contextes phonologiques. Dans ces contextes, nous utilisons les données connues pour calculer la répartition statistique des formes existantes entre les RTA en compétition. Cette distribution servira pour évaluer la part de ces règles dans la génération des formes fléchies.

Par exemple, dans le cas de la paire pr3sg⟶pr3pl, parmi toutes les RTA, la forme d'entrée fini (pr3sg) est compatible avec exactement l'ensemble des règles en (4) et nous savons d'après notre lexique-échantillon que la forme correspondante au pr3pl est finis (RTA 4b) tandis que ʁəni qui est compatible avec exactement le même ensemble de règles a pour forme du pr3pl ʁəni (RTA 4a). Ces deux formes fini et ʁəni sont prises en charge par le même ensemble des règles pouvant s'appliquer à leur configuration phonologique. Nous appelons ce type d'ensemble, une classe de compétition.

(4)
a) Ø⟶Ø/ X – (ʁəni⟶ʁəni, fini⟶*fini)
b) Ø⟶s / X[iɛea] – (ʁəni⟶*ʁənis, fini⟶finis)
c) Ø⟶t / X[ptbdfsvzmnʁɛəɒaɔɛ̃œ̃ɔ̃ɑ̃][jʁwɥiyɛeøœəauɔoɛ̃œ̃ɔ̃ɑ̃] – (ʁəni⟶*ʁənit, fini⟶*finit)

Dans la classe de compétition en (4), notre lexique-échantillon ne contient que des paires qui utilisent les RTA 4a et 4b. La distribution de l'ensemble des paires en question donne la répartition en (5).

(5)
a) Ø⟶Ø 20.69%
b) Ø⟶s 79.31%
c) Ø⟶t 0%

Ce type de répartition constaté dans le lexique-échantillon servira de base pour la phase de remplissage des paradigmes. La RTA 4c n'est jamais utilisée pour la classe de compétition en (4) mais son existence est justifiée indépendamment par la paire sɔʁ⟶sɔʁt qui n'entre pas dans cette classe de compétition du fait du contexte phonologique de 4b qui impose une finale vocalique (iɛea) pour les formes qui utilisent cette classe.

Pour ce travail, nous calculons non seulement les analogies et les classes de compétition entre toutes les cases du paradigmes comme Bonami & Boyé (2014), mais nous incluons en plus les analogies et les classes de compétition pour chaque case vers elle-même afin de capter les informations disponibles dans le lexique-échantillon sur la surabondance et de pouvoir inférer des surabondances systématiques comme celles de BALAYER, PAYER, ABOYER, …

9. voir §2.2.2, exemple (4)

Avec la constitution des lexiques-échantillons, l'extraction des RTA et des classes de compétition, l'étape de représentation des connaissances est achevée et nous disposons des données nécessaires au remplissage des paradigmes.

3 Le remplissage des paradigmes

Le remplissage des paradigmes procède en deux temps directement reliés avec l'observation fondamentale répétée ci-dessous :

(1) un paradigme flexionnel est un ensemble de formes toutes mutuellement reliées par des analogies

Dans l'idée que les formes fléchies sont reliées par des analogies, on commence par une génération analogique de toutes les formes possibles via les RTA et leurs classes de compétition à partir des formes présentes dans l'échantillon. Ceci crée des paradigmes massivement surabondants qui contiennent toutes les formes « imaginables » à partir des formes connues. Dans cette première étape, on a mobilisé, pour chaque lexème, les connaissances flexionnelles qui concernent l'appariement de ses formes connues avec celles de toutes les cases du paradigmes mais pas celles concernant les cases vides de l'échantillon. Pour mobiliser *toutes* les connaissances flexionnelles, pour tous les lexèmes, on applique ce procédé de génération une seconde fois, sur l'ensemble des formes générées à la première étape, pour obtenir toutes les formes « imaginables » au travers de l'ensemble de toutes les RTA et toutes les classes de compétition.

Avec ces paradigmes massivement surabondants, pour trouver des ensembles de formes cohérents, on utilise de nouveau l'observation (1). Un paradigme flexionnel étant un ensemble de formes toutes mutuellement reliées (clique [10]), on extrait chaque paradigme optimal en sélectionnant une clique triplement maximale de formes :

— maximale en tant que clique [11]

— maximale en couverture sur les cases du paradigmes

— maximale en nombre de formes dans la clique

3.1 Formes

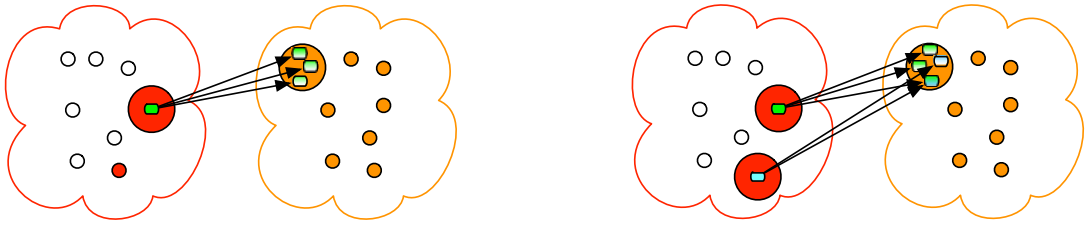
Pour la génération analogique des formes, nous parcourons le lexique-échantillon et nous créons toutes les formes analogiques potentielles dans toutes les cases du paradigme à partir des connaissances lexicales existantes.

Si la forme de départ appartient à une classe de compétition connue, elle crée des formes dans la case d'arrivée en proportion de la répartition calculée précédemment. Si elle n'appartient pas à une classe de compétition présente dans le lexique-échantillon, la case d'arrivée est remplie par l'ensemble des RTA de la classe en question avec une équirépartition. En revanche, si la forme ne correspond à aucune RTA, elle ne fournit aucune forme pour la case correspondante. Cette génération est illustrée dans la figure 2 à gauche où la forme de départ dans la case rouge agrandie appartient à une classe de compétition contenant 3 RTA productives [12] qui créent 3 nouvelles formes dans la case d'arrivée orange.

10. Dans un graphe non-orienté, une clique est un ensemble de nœuds tous mutuellement reliés deux à deux.

11. Une clique maximale est une clique qui ne se trouve pas inclue à l'intérieur d'une clique plus grande.

12. Les RTA productives d'une classe de compétition sont celles associées à un pourcentage non-nul pour les formes de sorties.

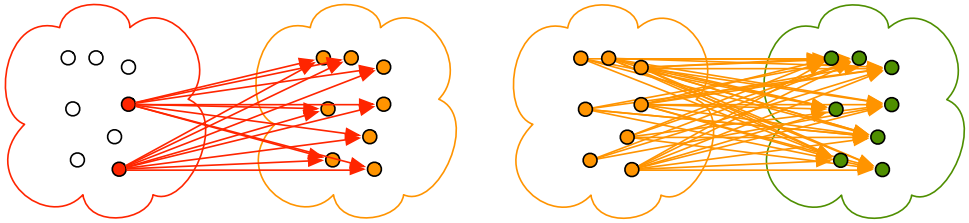


Génération Consolidation

FIGURE 2 – Remplissage et Répartition

Ce processus se répète pour chaque forme de chaque case du lexique-échantillon en direction de toutes les cases du paradigmes. Chaque case d'arrivée reçoit donc un certain nombre de formes et nous calculons la répartition de ces formes dans la case en balançant l'ensemble des résultats comme dans une élection. Chaque case de départ représente un bureau de vote pour la case d'arrivée, et le score de chaque forme de la case d'arrivée se calcule par consolidation entre les bureaux de vote, comme illustré figure 2 à droite où les formes des cases de départ (en rouge à gauche) cumulent leurs votes sur le candidat en bas au centre de la case d'arrivée tandis que le reste de leurs votes concernent des formes différentes.

À la fin de cette première vague de génération, seules les cases du lexique-échantillon qui contenaient des formes lexicales ont participé et en conséquence, seules les connaissances flexionnelles relatives à ces cases ont été mobilisées pour chaque lexème. Le but de la deuxième vague est d'utiliser toutes les connaissances flexionnelles mobilisables dans la génération des formes.



Première vague Deuxième vague

FIGURE 3 – Le double remplissage

À partir du nouveau lexique dérivé de l'échantillon à l'issue de la première vague, nous reproduisons le même processus en cherchant spécifiquement les liens qui s'établissent entre les formes générées par le premier tour pour trouver des relations mutuelles entre formes qui nous permettront de constituer des paradigmes.

3.2 Paradigmes

Après les deux vagues de génération, toutes les connaissances lexicales et flexionnelles mobilisables ont été utilisées. Il s'agit maintenant de filtrer les résultats obtenus pour faire émerger des paradigmes flexionnels du chaos surabondant créé précédemment. De nouveau, nous nous appuyons sur le concept de paradigme flexionnel en tant que réseau relationnel pour sélectionner les bons ensembles de formes.

Pour chaque lexème, nous cherchons les cliques de formes fléchies au sein du réseau créé lors de la deuxième vague de génération. Parmi toutes les cliques trouvées, nous éliminons celles qui sont contenues dans des cliques plus grandes et ne conservons que les cliques maximales.

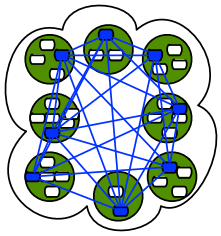


FIGURE 4 – une clique-paradigme

Pour ces cliques maximales, c'est l'ampleur de la couverture des cases du paradigme qui va guider le choix, le remplissage du paradigme est fonction du nombre de cases contenant des formes fléchies identifiées.

Chaque clique à couverture maximale reçoit une évaluation qui est fonction du support mutuel des formes entre elles. Seules les cliques recevant la note maximale sont conservées. Si plusieurs cliques ont la note maximale, elles sont fusionnées [13].

Ce processus aboutit à une clique-paradigme (cf. Fig. 4) optimale par lexème présent dans le lexique-échantillon qui représente la réponse à la question du remplissage des paradigmes. C'est ce résultat qui fait l'objet de l'évaluation élaborée dans la section suivante.

4 Analyse des résultats

Dans l'état actuel, le remplissage des paradigmes à partir d'un lexique-échantillon de 15 000 formes donne des résultats mitigés. D'un côté, même avec un échantillon très limité de 15.000 formes (4,6%), on obtient plus de 62.000 nouvelles formes, avec une précision encourageante (88,8%) mais le rappel [14] est très faible (46,6%). De nombreuses paires de cases du paradigmes ne correspondent qu'à une ou zéro formes (en jaune sur la figure 5) et ne permettent soit aucune RTA soit une RTA circonscrite à un verbe spécifique. La présence des barres jaunes limite la taille des paradigmes-cliques constituables et réduit d'autant le rappel. Avec un échantillon de 15 000 formes, les paires de cases qui comportent plus de 100 verbes (en vert) sont très limitées, et l'essentiel de la surface est distribuée entre des quantités de verbes entre 2 et 25 (en orange) et entre 25 et 100 (en bleu). Les barres jaunes et oranges correspondent en majorité à des formes du passé simple et du subjonctif imparfait qui sont de fait à la fois rares et peu maîtrisées même par les locuteurs natifs dont on peut légitimement se demander si le rappel est nécessaire. À ce niveau d'échantillonnage, la surface verte qui correspond aux paires bien étayées (plus de 100 verbes) est encore assez faible mais elle croît nettement avec la taille des échantillons comme le montre la figure 6.

Bien évidemment avec des lexiques-échantillons plus grands, les résultats sont nettement supérieurs mais le rappel reste faible avec trop peu de créations de nouvelles formes jusqu'à des taille d'échantillons aux alentours de 20% (cf. table 2). Pour pallier les problèmes de rappel, on peut envisager d'utiliser des RTA hors contexte pour les paires faiblement étayées (zones jaune et orange) pour augmenter la capacité générative du système dans les contextes trop réduits et n'utiliser des RTA avec la génération minimale que pour des paires bien étayées (zones bleue et verte).

13. Dans tous les cas observés jusqu'à présent, ces cliques formaient des sous-variantes d'un paradigme plus grand. À toutes fins utiles, la fusion est limitée aux cliques unifiables.

14. La précision mesure le nombre de précisions correctes sur le nombre total de prédictions, le rappel mesure le nombre de prédictions correctes sur l'ensemble des prédictions attendues.

15. Le tableau dans cette figure n'a pas été symétrisé et les données se limitent à la partie inférieure gauche.

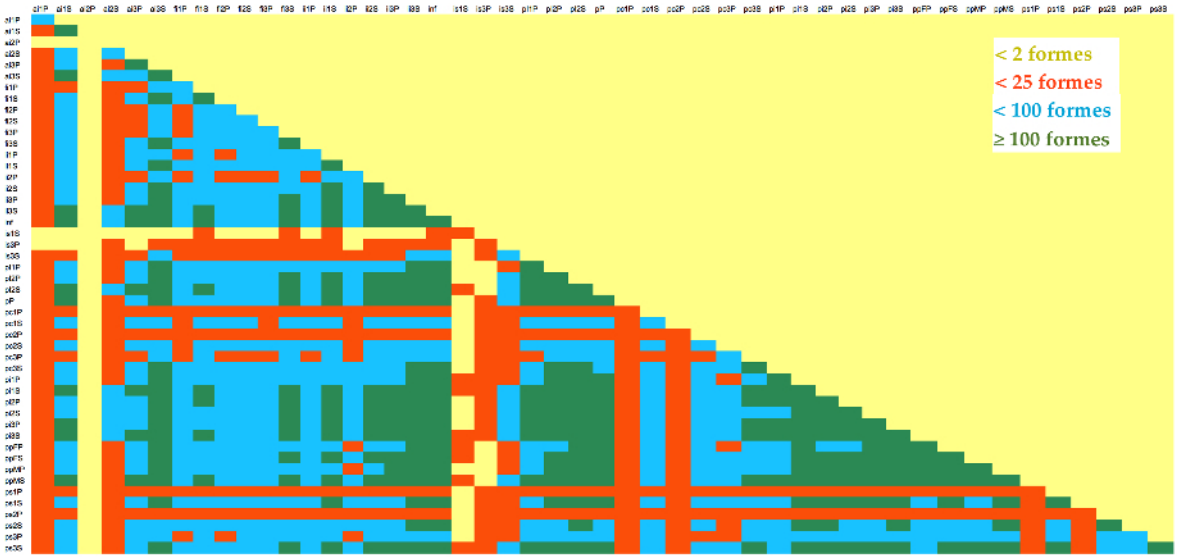
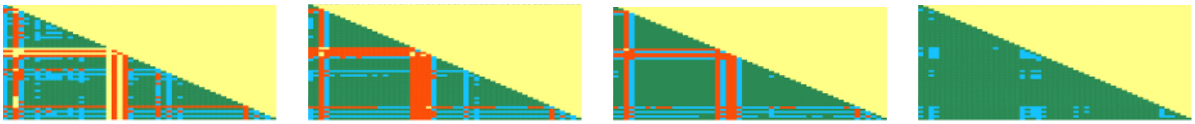


FIGURE 5 – Distribution du nombre de verbes disponibles par paires de formes [15]



30 000 formes 45 000 formes 60 000 formes 75 000 formes

FIGURE 6 – Évolution de la distribution en fonction de la taille des lexiques-échantillons

D'autre part, l'intégration de traits inhérents permettraient d'améliorer la définition des contextes des RTA, par exemple, en ajoutant une information lexicale pour générer de façon appropriée les participes passés féminins et masculins pluriels. Pour une analyse efficace de ces formes de participes passés, on ajouterait une information sur le verbe qui permettrait d'associer l'existence d'une forme unique avec la sélection de l'auxiliaire AVOIR pour les verbes ne prenant pas de complément d'objet direct. Cette information pourrait être utilisée lors de la formation des classes de compétition pour intervenir dans le calcul de la distribution.

5 Conclusion

Cette première modélisation de la conjugaison du français montre qu'une analyse surfaciste est envisageable. Comme on a pu le constater, surfaciste prend ici deux sens : (i) ne pas recourir à des analyses abstraites (radicaux, exposants), (ii) utiliser une étendue partielle du lexique et des connaissances flexionnelles, la question du remplissage des para-

Échantillon	Surface	Créations	Précision	Rappel	F-Score
15 000	4,6%	62 816	88,8%	46,6%	61,1%
30 000	9,1%	110 861	92,3%	61,0%	73,5%
45 000	13,7%	133 245	93,5%	66,6%	77,8%
60 000	18,3%	147 135	94,5%	71,1%	81,2%
75 000	22,9%	174 739	97,5%	86,3%	91,5%

TABLE 2 – Précisions et Rappels pour les différentes tailles de lexique-échantillon

digmes émergeant précisément du fait de la couverture partielle de cette surface. Les données utilisées sont accessibles aux locuteurs natifs contrairement aux classes flexionnelles ou au lexique complet nécessaires aux autres approches (p.e. Stump & Finkel, 2013; Bonami & Beniamine, 2015).

Dans d'autres contextes, comme la flexion nominale des langues slaves, il est envisageable que le lexique-échantillon contienne de nombreux paradigmes complets et qu'en conséquence l'approche par classes flexionnelles ne rencontre pas les mêmes problèmes d'accessibilité des données qu'ici. Mais l'approche que nous proposons peut être étendue à ce type de données tandis que l'inverse semble difficile.

Le modèle proposé repose sur des analogies entre toutes les formes des lexèmes qui permettent de remplir les paradigmes sans jamais référer à des classes flexionnelles mais de par son organisation, il ne peut rendre compte pour l'instant de la défectivité. La question posée est simple : quel type de connaissance lexicale et flexionnelle correspond à ce phénomène ? Dans l'état actuel, le modèle permet de capter les surabondances systématiques et isolées mais aucune défectivité.

Nous n'avons abordé ici que la question du remplissage des paradigmes mais une autre question tout aussi intéressante se pose aux locuteurs dans cette approche, celle de l'identification des formes inconnues de lexèmes connus. Avec l'approche proposée ici, on peut imaginer un processus simple permettant de rapprocher une forme extérieure au lexique-échantillon de celles d'un verbe connu : on effectue le remplissage du paradigme à partir de la forme extérieure et on compare ce paradigme avec ceux obtenus à partir des formes intérieures. Si la forme extérieure correspond à un verbe du lexique-échantillon, on s'attend à ce que son paradigme rempli soit le même que celui du verbe intérieur en question.

Références

Ackerman, Farrell, James P Blevins & Robert Malouf. 2009. Parts and wholes : Implicative patterns in inflectional paradigms. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar : Form and acquisition*, 54–82. Oxford Scholarship Online.

Albright, Adam. 2002. *The identification of bases in morphological paradigms* : UCLA dissertation.

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in english past tenses : A computational/experimental study. *Cognition* 90. 119–161.

Anderson, Stephen R. 1992. *A-Morphous Morphology* (Cambridge Studies in Linguistics 62). Cambridge University Press.

Arrivé, Michel. 1997. *La conjugaison pour tous*. Bescherelle. Hatier.

Beniamine, Sarah & Benoît Sagot. 2015. Segmentation strategies for inflection class inference. In *Décembrettes 9, Colloque international de morphologie*, Toulouse, France : Université de Toulouse. https ://hal.inria.fr/hal-01190524.

Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42. 531–573.

Bonami, Olivier & Sarah Beniamine. 2015. Implicative structure and joint predictiveness. In Vito Pirrelli, Claudia Marzi & Marcello Ferro (eds.), *Word structure and word usage : Proceedings of the networds final conference*, .

Bonami, Olivier & Gilles Boyé. 2014. De formes en thèmes. *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux* 17–45.

Bonami, Olivier & Ana R. Luís. 2014. Sur la morphologie implicative dans la conjugaison du portugais : une étude quantitative. *Mémoires de la Société de Linguistique de Paris* 22. 111–151.

Brown, Dunstan & Andrew Hippisley. 2012. *Network morphology : A defaults-based theory of word structure*, vol. 133 Cambridge Studies in Linguistics. Cambridge University Press.

de Calmès, Martine & Guy Pérennou. 1998. Bdlex : a lexicon for spoken and written french. In *1st international conference on language resources and evaluation*, 1129–1136. Grenade : ELRA.

Corbett, Greville G & Norman M Fraser. 1993. Network morphology : a datr account of russian nominal inflection. *Journal of Linguistics* 29(01). 113–142.

Gaume, Bruno, Emmanuel Navarro, Yann Desalle & Benoît Gaillard. 2014. Mesurer la similarité structurelle entre réseaux lexicaux. In *21ème taln*, 30–39.

Kilani-Schoch, Marianne & Wolfgang U. Dressler. 2005. *Morphologie naturelle et flexion du verbe français.* Tübingen : Gunter Narr Verlag.

New, Boris, C. Pallier, L. Ferrand & R. Matos. 2001. Une base de données lexicales du français contemporain sur internet : Lexique. *L'Année Psychologique* 101. 447–462.

Shannon, Claude. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27. 379–423, 623–656.

Stump, Gregory & Raphael A Finkel. 2013. *Morphological typology : From word to paradigm*, vol. 138. Cambridge University Press.

Stump, Gregory T. 2001. *Inflectional morphology.* Cambridge University Press.