Glawinette : construction d'un lexique morphologique dérivationnel du français

Nabil Hathout

CLLE, CNRS & Université Toulouse Jean Jaurès

Thématiques actuelles de la recherche en TAL
15 novembre 2021

Lexiques paradigmatiques

Extraire les couples de mots morphologiquement apparentés des définitions de GLAWI

BAP: Identifier le patrons d'alternance global de chaque couple de mots

Identifier les régularités formelles dans les séries de mots et les séries de couples

FAP: Identifier les exposants « d'usage »

Contexte

En collaboration avec :

Basilio Calderone	(CLLE)
Franck Sajous	(CLLE)
Fiammetta Namer	(ATILF)

Objectif à long terme :

Construire un lexique dérivationnel paradigmatique du français, de l'anglais et de l'italien =
Construire un **Bescherelle de la dérivation** qui fournit :

- le paradigme dérivationnel de chaque mot
  ($\sim$ numéro de la table de conjugaison)
- la liste des autres mots de sa famille morphologique
  ($\sim$ table de conjugaison elle-même).

Les paradigmes en flexion

La morphologie flexionnelle est paradigmatique. Les paradigmes flexionnels s'appellent des **classes flexionnelles**.

Les formes fléchies des verbes sont classiquement décrites au moyen de **tables de conjugaison** (Bescherelle).

Classe flexionnelle de *laver*

	Vmip1s-	Vmip2s-	...	Vmif1p-	Vmif2p-	...	Vmcp3s-	Vmcp3p-	...
LAVER	lave	laves	...	lavons	lavez	...	laverait	laveraient	...
CASSER	casse	casses	...	cassons	cassez	...	casserait	casseraient	...
ÉCLAIRER	éclaire	éclaires	...	éclairons	éclairez	...	éclairerait	éclaireraient	...
SALUER	salue	salues	...	saluons	saluez	...	saluerait	salueraient	...

Les paradigmes flexionnels du français peuvent être représentés dans des tables qui contiennent : 51 colonnes pour un verbe,
4 colonnes pour un adjectif,
2 colonnes pour un nom,
1 colonne pour un adverbe.

Les paradigmes en flexion

Classe flexionnelle de *laver*

	Vmip1s-	Vmip2s-	...	Vmif1p-	Vmif2p-	...	Vmcp3s-	Vmcp3p-	...
LAVER	lave	laves	...	lavons	lavez	...	laverait	laveraient	...
CASSER	casse	casses	...	cassons	cassez	...	casserait	casseraient	...
ÉCLAIRER	éclaire	éclaires	...	éclairons	éclairez	...	éclairerait	éclaireraient	...
SALUER	salue	salues	...	saluons	saluez	...	saluerait	salueraient	...

- Les lignes décrivent l'ensemble des formes du lexème.
- Les colonnes décrivent les formes qui réalisent les mêmes combinaisons de traits morphosyntaxiques des lexèmes d'une même classe flexionnelle.

Les paradigmes en dérivation

La description paradigmatique de la morphologie dérivationnelle permet de rendre de compte simplement de nombreux phénomènes non canoniques :

- relations indirectes (*prédateur* $\leftrightarrow$ *prédation*) ;
- parasynthèses (*banque* $\rightarrow$ *interbancaire*) ;
- rétro-formations (*to babysit* $\leftarrow$ *baby-sitter*) ;
- syncrétismes
  (*français* = 'gentilé de France' = 'langue parlée en France') ;
- surabondances (*cerisaie* / *ceriseraie* = 'verger de cerisiers') ; etc.

Bochner (1993); Bauer (1997); Štekauer (2014); Antoniova & Štekauer (2015); Blevins (2016); Hathout & Namer (2018a,b); Bonami & Strnadová (2019); Hathout & Namer (2019); Namer & Hathout (2020)

Les paradigmes en dérivation

Transposition à la morphologie dérivationnelle des tables de conjugaison :

Paradigme dérivationnel (partiel)

LAVER	LAVAGE	LAVEUR	LAVEUSE	LAVABLE
CASSER	CASSAGE	CASSEUR	CASSEUSE	CASSABLE
ÉCLAIRER	ÉCLAIRAGE	ÉCLAIREUR	ÉCLAIREUSE	ÉCLAIRABLE
SOUDER	SOUDAGE	SOUDEUR	SOUDEUSE	SOUDABLE

Les paradigmes en dérivation

Paradigme dérivationnel (partiel)

LAVER	LAVAGE	LAVEUR	LAVEUSE	LAVABLE
CASSER	CASSAGE	CASSEUR	CASSEUSE	CASSABLE
ÉCLAIRER	ÉCLAIRAGE	ÉCLAIREUR	ÉCLAIREUSE	ÉCLAIRABLE
SOUDER	SOUDAGE	SOUDEUR	SOUDEUSE	SOUDABLE

- Les lignes décrivent des **familles morphologiques** = ensembles de lexèmes morphologiquement apparentés. (Roché, 2009; Hathout, 2009c, 2011b)
- Les colonnes décrivent des **séries morphologiques de lexèmes** = ensembles de lexèmes qui ont les mêmes contrastes de forme et de sens avec les autres membres de leurs familles morphologiques.
- Les couples de lignes décrivent des **séries morphologiques de couples de lexèmes** = des ensembles de couples de lexèmes qui présentent les mêmes contrastes de forme et de sens = paradigme binaire.

Glawinette, Glawitina, Englawinette

Glawinette = lexique dérivationnel extrait du dictionnaire électronique du **français** GLAWI (Sajous & Hathout, 2015; Hathout & Sajous, 2016; Hathout et al., 2020)

Glawitina = lexique dérivationnel extrait du dictionnaire électronique de l'**italien** GLAW-IT (Calderone et al., 2016)

Englawinette = lexique dérivationnel extrait du dictionnaire électronique de l'**anglais** ENGLAWI (Sajous et al., 2020)
En cours de finalisation.

Démonette

Glawinette est destiné à alimenter la base de données morphologique Démonette (Hathout & Namer, 2014a,b, 2016) en cours en construction dans le cadre du projet **ANR Demonext** (Namer et al., 2019).

Glawinette, Glawitina, Englawinette

Les trois lexiques ont exactement la même structure et sont construits en utilisant les mêmes programmes.
Ils fournissent les structures nécessaires à la construction de paradigmes dérivationnels :

- des familles morphologiques composées de **couples de lexèmes morphologiquement apparentés**.
- la description des contrastes de forme (et de sens) entre les **couples lexèmes morphologiquement apparentés**.

Les contrastes de forme sont considérés comme les reflets des contrastes de sens (Hathout, 2008, 2009b,a, 2011a).

Glawinette, Glawitina, Englawinette

Les familles morphologiques sont décrites sous forme de graphes de lexèmes connectés par des relations dérivationnelles.

Famille morphologique

prince=N:princesse=N
prince=N:princier=A
prince=N:princillon=N
prince=N:princiser=V
princesse=N:prince=N
princier=A:prince=N
princier=A:princièrement=R
princillon=N:prince=N
princiser=V:prince=N
princièrement=R:princier=A

Glawinette, Glawitina, Englawinette

Les séries de couples sont des ensembles de couples qui présentent les mêmes contrastes de forme (et de sens).

Série de couples de lexèmes

`^(.+)eur$=N`	`^(.+)ion$=N`
acteur	action
animateur	animation
classificateur	classification
colonisateur	colonisation
directeur	direction
décentralisateur	décentralisation
dépresseur	dépression
éditeur	édition
expositeur	exposition
formateur	formation
réacteur	réaction
réviseur	révision

Glawinette, Glawitina, Englawinette

Glawinette

77 682	lexèmes
144 028	couples de lexèmes morphologiquement apparentés
15 843	familles dérivationnelles
5 384	séries de relations dérivationnelles

Extrait de la famille de *serrer*

desserrer=V:serrer=V resserrer=V:serrer=V
resserreur=N:resserrer=V réenserrer=V:enserrer=V
serrage=N:serrer=V serre=N:enserrer=V **serre=N:serrer=V**
serre=N:serriste=N **serrement=N:serrer=V** serrer=V:desserrer=V
serrer=V:resserrer=V serrer=V:serrage=N serrer=V:serre=N
serrer=V:serrement=N serrer=V:serrure=N serrer=V:serré=A
serriste=N:serre=N serrure=N:serrer=V serrure=N:serrurerie=N
serrure=N:serrurier=N serrurerie=N:serrure=N

Caractérisation des contrastes des couples de lexèmes morphologiquement apparentés

serrer=V	serrement=N	^(.+)er$	V	^(.+)ement$	N
serrer=V	serrure=N	^(.+)er$	V	^(.+)ure$	N
serrure=N	serrurier=N	^(.+)e$	N	^(.+)ier$	N
serrurier=N	serrure=N	^(.+)ier$	N	^(.+)e$	N
serrurerie=N	serrure=N	**^(.+)erie$**	N	^(.+)e$	N
serrurerie=N	serrurier=N	**^(.+)erie$**	N	**^(.+)ier$**	N
serrure=N	serrurerie=N	^(.+)e$	N	^(.+)erie$	N
desserre=N	desserrer=V	^(.+)e$	N	^(.+)er$	V
desserrer=V	desserrage=N	^(.+)er$	V	^(.+)age$	N
desserrage=N	desserrer=V	**^(.+)age$**	N	^(.+)er$	V
indesserrable=A	desserrer=V	**^in(.+)able$**	A	^(.+)er$	V
desserroir=N	desserrer=V	**^(.+)oir$**	N	^(.+)er$	V
enserrer=V	enserrement=N	^(.+)er$	V	^(.+)ement$	N
enserrer=V	renserrer=V	^(.+)er$	V	^r(.+)er$	V
enserrer=V	serre=N	**^en(.+)er$**	V	^(.+)e$	N
resserrer=V	resserrement=N	^(.+)er$	V	^(.+)ement$	N
resserrement=N	resserrer=V	**^(.+)ement$**	N	^(.+)er$	V
réenserrer=V	enserrer=V	^ré(.+)er$	V	^(.+)er$	V

Famille de *battere* dans Glawitina

abbattere=V:abbattibile=A abbattere=V:abbattimento=N
abbattere=V:abbattitore=A abbattere=V:abbattitore=N
abbattere=V:abbattuta=N abbattere=V:battere=V
abbattibile=A:abbattere=V abbattimento=N:abbattere=V
abbattimento=N:abbattitore=N abbattitore=A:abbattere=V
abbattitore=N:abbattere=V abbattitore=N:abbattimento=N
abbattitore=N:abbattitrice=N abbattitrice=N:abbattitore=N
abbattuta=N:abbattere=V battente=N:battere=V
battere=V:abbattere=V battere=V:battente=N battere=V:battitura=N
battere=V:battuta=N battere=V:imbattibile=A battere=V:ribattere=V
battere=V:sbattere=V battitura=N:battere=V battuta=N:battere=V
imbattibile=A:battere=V ribattere=V:battere=V sbattere=V:battere=V

Caractérisation des couples (Glawitina)

abbattere=V	battere=V	^abba(.+)re$	V	^ba(.+)re$	V
battere=V	sbattere=V	^(.+)re$	V	^s(.+)re$	V
battere=V	imbattibile=A	^(.+)ere$	V	^im(.+)ibile$	A
battere=V	ribattere=V	^(.+)re$	V	^ri(.+)re$	V
abbattitrice=N	abbattitore=N	^(.+)trice$	N	^(.+)tore$	N
abbattere=V	abbattimento=N	^(.+)ere$	V	^(.+)imento$	N
abbattere=V	abbattibile=A	^(.+)ere$	V	^(.+)ibile$	A
battere=V	battitura=N	^(.+)ere$	V	^(.+)itura$	N
ribattere=V	battere=V	^ri(.+)re$	V	^(.+)re$	V
abbattere=V	abbattitore=A	^(.+)ere$	V	^(.+)itore$	A
abbattitore=A	abbattere=V	^(.+)itore$	A	^(.+)ere$	V
abbattuta=N	abbattere=V	^(.+)uta$	N	^(.+)ere$	V
battitura=N	battere=V	^(.+)itura$	N	^(.+)ere$	V
battere=V	abbattere=V	^ba(.+)re$	V	^abba(.+)re$	V
battente=N	battere=V	^(.+)nte$	N	^(.+)re$	V
imbattibile=A	battere=V	^im(.+)ibile$	A	^(.+)ere$	V
abbattere=V	abbattitore=N	^(.+)ere$	V	^(.+)itore$	N
abbattimento=N	abbattitore=N	^(.+)timento$	N	^(.+)titore$	N

Famille de *scratch* dans Englawinette

backscratch=N:scratch=N backscratch=V:scratch=V
bescratch=V:scratch=V **catscratch=N:scratch=N cratch=V:scratch=V**
headscratcher=N:scratcher=N nonscratchable=A:scratch=V
nonscratchable=A:scratchable=A scratch=A:scratcher=N
scratch=A:scratchy=A scratch=N:scratchcard=N scratch=N:scratcher=N
scratch=N:scratchlike=A scratch=N:scratchmark=N
scratch=N:scratchpad=N scratch=N:scratchpaper=N scratch=V:cratch=V
scratch=V:scratchable=A scratch=V:scratchband=N scratch=V:scratching=N
scratch=V:scratchy=A scratchability=N:scratchable=A
scratchable=A:scratchability=N scratchable=A:unscratchable=A
scratchband=N:scratch=N scratchband=N:scratch=V
scratchcard=N:scratch=N scratched=A:scratching=N
scratcher=N:headscratcher=N scratcher=N:scratch=A
scratcher=N:scratch=N scratcher=N:scratch=V scratchie=N:scratch=N
scratchily=R:scratchy=A scratchiness=N:scratchsome=A
scratchiness=N:scratchy=A scratching=N:scratch=V
scratching=N:scratchsome=A scratching=V:scratch=N
scratchingly=R:scratching=N ...

Caractérisation des relations (Englawinette)

scratchy=A	scratch=N	^(.+)y$	A	^(.+)$	N
scratchy=A	scratch=V	^(.+)y$	A	^(.+)$	V
scratcher=N	scratch=N	^(.+)er$	N	^(.+)$	N
scratching=N	scratchingly=R	^(.+)$	N	^(.+)ly$	R
unscratchable=A	scratchable=A	^un(.+)$	A	^(.+)$	A
scratchable=A	nonscratchable=A	^(.+)$	A	^non(.+)$	A
headscratcher=N	scratcher=N	^head(.+)$	N	^(.+)$	N
scratch=N	scratchless=A	^(.+)$	N	^(.+)less$	A
scratchlike=A	scratch=N	^(.+)like$	A	^(.+)$	N
scratchpaper=N	scratch=N	^(.+)paper$	N	^(.+)$	N
scratching=N	scratched=A	^(.+)ing$	N	^(.+)ed$	A
scratchy=A	scratchiness=N	^(.+)y$	A	^(.+)iness$	N
scratch=V	nonscratchable=A	^(.+)$	V	^non(.+)able$	A
scratching=N	scratchsome=A	^(.+)ing$	N	^(.+)some$	A
scratchsome=A	scratchiness=N	^(.+)some$	A	^(.+)iness$	N
scratch=A	scratcher=N	^(.+)$	A	^(.+)er$	N
scratch=V	backscratch=V	^(.+)$	V	^back(.+)$	V
scratchable=A	scratch=V	^(.+)able$	A	^(.+)$	V

Fondements de la construction de Glawinette

1. Les définitions (morphologiques) fournissent des couples de lexèmes en relation sémantique. Ces couples sont très bruités.
2. L'analogie permet d'identifier des régularités formelles des couples de lexèmes et de les filtrer.
3. Les régularités formelles des couples lexèmes déterminent des séries de couples.
4. L'analogie permet d'identifier les régularités formelles des séries de lexèmes. Les régularités des séries de lexèmes peuvent être alignées pour caractériser plus finement les séries de couples.
5. Les couples réguliers forment un graphe dérivationnel dont les composantes connexes sont les familles morphologiques.
6. Les couples peuvent caractériser au moyen de régularités fines composées d'exposants « usuels » en minimisant la longueur de description du lexique.

Lexiques paradigmatiques

Extraire les couples de mots morphologiquement apparentés des définitions de GLAWI

BAP: Identifier le patrons d'alternance global de chaque couple de mots

Identifier les régularités formelles dans les séries de mots et les séries de couples

FAP: Identifier les exposants « d'usage »

Définitions morphologiques

- Une définition morphologique décrit le sens d'un mot construit relativement à un autre mot de sa famille dérivationnelle (Martin, 1992).
- Une grande partie des lexèmes morphologiquement construits sont définis par des définitions morphologiques.
- Le mot de la famille n'est pas toujours la base du mot construit

clocheton = petit bâtiment en forme de **clocher**, de tourelle, dont on orne les angles ou le sommet d'une construction

glaçon = morceau de **glace**

développement = action de **développer**, de se **développer** ou résultat de cette action, au propre et au figuré

productivisme = doctrine selon laquelle la **production** est un objectif premier, système qui prône le sacrifice de toute autre considération pour maximiser la **productivité**

Définitions morphologiques

Nous ne savons pas identifier les définitions morphologiques.

Nous extrayons de GLAWI **tous les couples de mots** $(w1, w2)$ où w_1 est un défini et w_2 un mot qui apparaît dans son définissant, tels que :

- w_1 et w_2 sont de catégories majeures (N, V, Adj, Adv) ;
- ni w_1 ni w_2 ne soient des *stop-words* (antidictionnaire NLTK).

Nous utilisons les définitions analysées en dépendances pour déterminer les catégories grammaticales de w_1 et w_2.

2 184 847 couples de mots ont été extraits de GLAWI.

Identifier des régularités formelles au moyen de l'analogie

- Les contrastes formels des couples de lexèmes morphologiquement apparentés présentent des régularités.
- Les couples de mots qui ont les mêmes contrastes formels forment des **analogies**

développement : développer : :classement : classer forment une analogie formelle (Lepage, 2004a,b; Stroppa & Yvon, 2005). Le contraste formel peut être décrit :

- comme une substitution

  `/^(.+)er$/^\1ement$/`

- par un patron composé de deux expressions régulières

  `^(.+)er$/^(.+)ement$`

  où `(.+)` représente une sous-chaîne de caractères identique dans les deux lemmes.

Identifier des régularités formelles au moyen de l'analogie

Nous utilisons la méthode proposée par Lepage (1998, 2004b) pour identifier les quadruplets analogiques :

- si $A{:}B{::}C{:}D$ alors la distance de Levenshtein entre A et B est égale à la distance de Levenshtein entre C et D
- si $A{:}B{::}C{:}D$ alors, pour chaque caractère a de l'alphabet, $|A|_a - |B|_a = |C|_a - |D|_a$ où $|X|_a$ représente le nombre d'occurrences de a dans X.

On associe à chaque couple de mots (A, B) une signature (Hathout, 2002, 2003, 2005) :

$$\sigma(A, B) = (d(A, B), |A|_{a_1} - |B|_{a_1}, \ldots, |A|_{a_n} - |B|_{a_n})$$

où $d(A, B)$ est la distance de Levenshtein entre A et B et où $\{a_1, \ldots, a_n\}$ est l'alphabet du langage.

Identifier des régularités formelles au moyen de l'analogie

Deux couples de mots qui ont la même signature forment une analogie, même s'il existe quelques exceptions.

Les signatures sont calculées en utilisant la bibliothèque `fast_distance` de Yves Lepage et le module Python `collections`.

Identifier des régularités formelles au moyen de l'analogie

Les relations de forme et de sens régulières sont des relations morphologiques dérivationnelles;
mais on ne sait pas si les relations sémantiques entre les couples de mots extraits des définitions de GLAWI sont régulières.

On sait que les relations formelles entre les couples de mots qui ont les mêmes signatures sont régulières (**développement : développer** et **classement : classer**).

Les relations formelles les moins fréquentes sont généralement accidentelles (**clocheton :angle**).

$\Rightarrow$ Nous excluons les couples de lexèmes dont la signature a une fréquence inférieure à 5 dans l'ensemble des couples extraits de GLAWI.

Lexiques paradigmatiques

Extraire les couples de mots morphologiquement apparentés des définitions de GLAWI

BAP: Identifier le patrons d'alternance global de chaque couple de mots

Identifier les régularités formelles dans les séries de mots et les séries de couples

FAP: Identifier les exposants « d'usage »

Caractérisation globale des régularités formelles

Les séries de couples peuvent être caractérisées au moyen de patrons d'alternance globaux (*Broad Alternation Pattern*, BAP).

Le BAP d'un couple de lexèmes est un couple d'expressions régulières qui décrit **la substitution la plus générale** qui permet de transformer le lemme du premier en celui du second. Les séquences `(.+)` représentent des sous-chaînes identiques dans les deux lemmes.

enserrer=V	**enserrement=N**	**^(.+)r$**	**^(.+)ment$**
serrer=V	**serrure=N**	**^(.+)e(.+)$**	**^(.+)u(.+)e$**

Caractérisation globale des régularités formelles

enserrer=V	**enserrement=N**	**^(.+)r$**	**^(.+)ment$**
serrer=V	**serrure=N**	**^(.+)e(.+)$**	**^(.+)u(.+)e$**

✔ Les BAP permettent de distinguer les séries de couples *-eur/-age* (**allumeur:allumage** ; ^(.+)(.+)ur$/^(.+)ag(.+)$) des couples *-ure/-age* (**doublure:doublage** ; ^(.+)ur(.+)$/^(.+)ag(.+)$).
Les couples des deux séries ont la même signature analogique.

✗ Les BAP ne sont pas linguistiquement motivés.

✗ Le BAP ^(.+)e(.+)$/^(.+)u(.+)e$ décrit **serrer:serrure** mais aussi **formel:formule**.

Lexiques paradigmatiques

Extraire les couples de mots morphologiquement apparentés des définitions de GLAWI

BAP: Identifier le patrons d'alternance global de chaque couple de mots

Identifier les régularités formelles dans les séries de mots et les séries de couples

FAP: Identifier les exposants « d'usage »

Séparer les couples *-eur*/*-age* et *-ure*/*-age*

`^(.+)eur$`	`^(.+)age$`
allumeur	allumage
atterrisseur	atterrissage
balayeur	balayage
carreleur	carrelage
épandeur	épandage

- Les lemmes de la 1re colonne finissent tous en *-eur* : `eur$`
- Les lemmes de la 2e colonne finissent tous en *-age* : `age$`

`^(.+)ure$`	`^(.+)age$`
doublure	doublage
boursouflure	boursouflage
rayure	rayage
épluchure	épluchage

- Les lemmes de la 1re colonne finissent tous en *-ure* : `ure$`
- Les lemmes de la 2e colonne finissent tous en *-age* : `age$`

Séparer les couples *-eur*/*-age* et *-ure*/*-age*

$\Rightarrow$ *allumeur*:*allumage* et *doublure*:*doublage* ont la même signature, mais il n'y a pas d'analogie entre *allumeur*:*allumage* et *doublure*:*doublage*.

Les couples d'une série analogique formelle qui instancient des patrons de séries de mots différents appartiennent à des séries dérivationnelles différentes.

Séparer les couples *-eur*/*-age* et *-ure*/*-age*

Méthode

1. Dans chaque série, on identifie tous les patrons de mots qui décrivent un sous-ensemble des mots de chaque colonne.
2. On constitue des **patrons de couples** en alignant les patrons de mots dont les parties variables (`.+`) instancient la même chaîne de caractères dans les couples de mots de la série.

Séparer les couples *-eur*/*-age* et *-ure*/*-age*

Patrons des couples *-eur*/*-age* et leur couverture

Patron *-eur*	Cov
`^(.+)$`	1.0
`^(.+)r$`	1.0
`^(.+)u(.+)$`	1.0
`^(.+)ur$`	1.0
`^(.+)e(.+)$`	1.0
`^(.+)eu(.+)$`	1.0
`^(.+)eur$`	1.0
`^a(.+)$`	0.4
`^a(.+)r$`	0.4
`^a(.+)u(.+)$`	0.4
`^a(.+)ur$`	0.4
`^a(.+)e(.+)$`	0.4
`^a(.+)eu(.+)$`	0.4
`^a(.+)eur$`	0.4

Patron *-age*	Cov
`^(.+)$`	1.0
`^(.+)e$`	1.0
`^(.+)g(.+)$`	1.0
`^(.+)ge$`	1.0
`^(.+)a(.+)$`	1.0
`^(.+)ag(.+)$`	1.0
`^(.+)age$`	1.0
`^a(.+)$`	0.4
`^a(.+)e$`	0.4
`^a(.+)g(.+)$`	0.4
`^a(.+)ge$`	0.4
`^a(.+)a(.+)$`	0.4
`^a(.+)ag(.+)$`	0.4
`^a(.+)age$`	0.4

Régularités formelles des séries de couples de lexèmes

1. On compare tous les couples de lexèmes de la série deux-à-deux.
2. On calcule pour chaque couple de couples (w_1, w_2) et (w_3, w_4) le `diff` entre w_1 et w_3 et entre w_2 et w_4 (Bernhard, 2010)
   - Les séquences inchangées correspondent aux exposants (e.g. `eur`, `age`)
   - Les séquences modifiées correspondent aux radicaux (e.g. `allum`, `atterriss`)
3. Le patron est étendu en ajoutant une partie de longueur identique des sous-chaînes initiales ou finales des radicaux. On obtient ainsi l'ensemble des patrons qui décrivent (w_1, w_2, w_3, w_4)
4. On conserve les couples font les patrons :
   - ne contiennent qu'un seul radical `(.+)` ;
   - décrivent au moins 5 couples;
   - couvrent au moins 10% de la série formelle.

Lexiques paradigmatiques

Extraire les couples de mots morphologiquement apparentés des définitions de GLAWI

BAP: Identifier le patrons d'alternance global de chaque couple de mots

Identifier les régularités formelles dans les séries de mots et les séries de couples

FAP: Identifier les exposants « d'usage »

Caractérisation fine des régularités formelles

Les couples retenus sont généralement décrits par plusieurs patrons.

Nous proposons une méthode qui permet de choisir le patron dont les exposants sont les plus « usuels »

		patrons	**select.**
verbaliser=V	verbalisation=N	`^(.+)er$:^(.+)ation$`	←
		`^(.+)iser$:^(.+)isation$`	
proverbial=A	proverbialement=R'	`^(.+)ial$:^(.+)ialement$`	
		`^(.+)al$:^(.+)alement$`	
		`^(.+)l$:^(.+)lement$`	
		`^(.+)$:^(.+)ement$`	←
féministe=A	féminisme=N	`^(.+)niste$:^(.+)nisme$`	
		`^(.+)iste$:^(.+)isme$`	←
		`^(.+)ste$:^(.+)sme$`	
sarkozysme=N	sarkozyste=N	`^(.+)ste$:^(.+)sme$`	←

Caractérisation fine des régularités formelles

Il s'agit de minimiser la longueur de description du lexique en sélectionnant les radicaux et les exposants qui se combinent le plus fréquemment.

- Lorsqu'un couple (w_1, w_2) a plusieurs patrons, on sélectionne celui dont les exposants sont les plus « pertinents »
  - = qui apparaissent dans les patrons de couples les plus « connectants » **au niveau de l'ensemble du lexique**
  - = qui ont le plus grand nombre d'instances dans les couples des séries morphologiques du lexique.
- La « pertinence » d'un patron de couples (P, Q) est estimée par le nombre des lexèmes qui sont contenus dans le lexique et qui sont connectés à l'un des lexèmes décrits par les patrons de mots P et Q.

Caractérisation fine des régularités formelles

Méthode

- Soit R l'ensemble des patrons des couples $\{(X_1, Y_1), (X_2, Y_2), \ldots\}$ de la série qui contient (w_1, w_2).
- Soit C l'ensemble des couples morphologiques du lexique.
- Pour chaque patron de lexèmes $Z \in \{X_1, X_2, ...\} \cup \{Y_1, Y_2, ...\}$, on calcule $|Z|$ = le nombre de lexèmes qui apparaissent dans un couple de C et qui sont y décrits par Z.
- On sélectionne le patron de couples $(P, Q) \in R$ qui maximise $|P| + |Q|$.

Références

Antoniova, Vesna & Pavol Štekauer. 2015. Derivational paradigms within selected conceptual fields – contrastive research. *Facta Universitatis, Series: Linguistics and Literature* 13(2). 61–75.

Bauer, Laurie. 1997. Derivational paradigms. In *Yearbook of morphology 1996*, 243–256. Springer.

Bernhard, Delphine. 2010. Apprentissage non supervisé de familles morphologiques: Comparaison de méthodes et aspects multilingues. *Traitement Automatique des Langues* 51(2). 11–39.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

Bochner, Harry. 1993. *Simplicity in generative morphology*. Berlin & New-York: Mouton de Gruyter.

Bonami, Olivier & Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2). 167–197.

Références

Calderone, Basilio, Franck Sajous & Nabil Hathout. 2016. GLAW-IT: A free large Italian dictionary encoded in a fine-grained XML format. In *Proceedings of the 49th annual meeting of the societas linguistica europaea (sle 2016)*, 43–45. Naples, Italy.

Hathout, Nabil. 2002. From WordNet to CELEX: Acquiring morphological links from dictionaries of synonyms. In *Proceedings of the third international conference on language resources and evaluation*, 1478–1484. Las Palmas de Gran Canaria: ELRA.

Hathout, Nabil. 2003. L'analogie, un moyen de croiser les contraintes et les paradigmes. Acquisition de connaissances morphologiques à partir de dictionnaires de synonymes. *Revue d'Intelligence Artificielle* 17(5-6). 923–934.

Hathout, Nabil. 2005. Exploiter la structure analogique du lexique construit : Une approche computationnelle. *Cahiers de lexicologie* 87(2). 5–28.

Hathout, Nabil. 2008. Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the coling workshop textgraphs-3*, 1–8. Manchester: ACL.

Références

Hathout, Nabil. 2009a. Acquisition morphologique à partir d'un dictionnaire informatisé. In *Actes de la 16e conférence sur le traitement automatique des langues naturelles (taln-2009)*, Senlis: ATALA.

Hathout, Nabil. 2009b. Acquisition of morphological families and derivational series from a machine readable dictionary. In Fabio Montermini, Gilles Boyé & Jesse Tseng (eds.), *Selected proceedings of the 6th décembrettes: Morphology in bordeaux*, Somerville, MA: Cascadilla Proceedings Project.

Hathout, Nabil. 2009c. *Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie*. Toulouse: Université de Toulouse 2 - Le Mirail Habilitation à diriger des recherches.

Hathout, Nabil. 2011a. Morphonette: a paradigm-based morphological network. *Lingue e linguaggio* 2011(2). 243–262.

Hathout, Nabil. 2011b. Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*. In Roché et al. (2011) 251–318.

Références

Hathout, Nabil & Fiammetta Namer. 2014a. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5). 125–168.

Hathout, Nabil & Fiammetta Namer. 2014b. La base lexicale Démonette : entre sémantique constructionnelle et morphologie dérivationnelle. In *Actes de la 21e conférence annuelle sur le traitement automatique des langues naturelles (taln-2014)*, 208–219. Marseille: ATALA.

Hathout, Nabil & Fiammetta Namer. 2016. Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for French. In *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016)*, Portorož, Slovenia.

Hathout, Nabil & Fiammetta Namer. 2018a. Defining paradigms in word formation: concepts, data and experiments. *Lingue e Linguaggio* 17(2). 151–154.

Références

Hathout, Nabil & Fiammetta Namer. 2018b. La parasynthèse à travers les modèles : des RCL au ParaDis. In Olivier Bonami, Gilles Boyé, Georgette Dal, Hélène Giraudo & Fiammetta Namer (eds.), *The lexeme in descriptive and theroretical morphology* Empirically Oriented Theoretical Morphology and Syntax, 365–399. Berlin: Language science Press. `http://langsci-press.org/catalog/book/165`.

Hathout, Nabil & Fiammetta Namer. 2019. Paradigms in word formation: what are we up to? *Morphology* 29(2). 153–165.

Hathout, Nabil & Franck Sajous. 2016. Wiktionnaire's Wikicode GLAWIfied: a workable French machine-readable dictionary. In *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016)*, Portorož, Slovenia.

Hathout, Nabil, Franck Sajous, Basilio Calderone & Fiammetta Namer. 2020. Glawinette: a linguistically motivated derivational description of French acquired from GLAWI. In *Proceedings of the twelfth international conference on language resources and evaluation (LREC 2020)*, 3870–3878. Marseille.

Références

Lepage, Yves. 1998. Solving analogies on words: An algorithm. In *Proceedings of the 36th annual meeting of the association for computational linguistics and of the 17th international conference on computational linguistics*, vol. 2, 728–735. Montréal.

Lepage, Yves. 2004a. Analogy and formal languages. *Electronic notes in theoretical computer science* 53. 180–191.

Lepage, Yves. 2004b. Lower and higher estimates of the number of true analogies between sentences contained in a large multilingual corpus. In *Proceedings of the 20th international conference on computational linguistics (COLING-2004)*, 736–742. Genève.

Martin, Robert. 1992. *Pour une logique du sens* Linguistique nouvelle. Paris: Presses universitaires de France.

Namer, Fiammetta, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout & Delphine Tribout. 2019. Demonette2 - une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de la 26^{e} conférence annuelle sur le traitement automatique des langues naturelles (taln 2019)*, 233–243. Toulouse.

Références

Namer, Fiammetta & Nabil Hathout. 2020. ParaDis and Démonette – from theory to resources for derivational paradigms. *The Prague Bulletin of Mathematical Linguistics* 114. 5–33.

Roché, Michel. 2009. Pour une morphologie *lexicale*. In *La morphologie lexicale est-elle possible ?*, vol. 17 Mémoires de la Société de Linguistique, Nouvelle Série, 65–87. Leuven: Éditions Peeters.

Roché, Michel, Gilles Boyé, Nabil Hathout, Stéphanie Lignon & Marc Plénat. 2011. *Des unités morphologiques au lexique*. Paris: Hermès Science-Lavoisier.

Sajous, Franck, Basilio Calderone & Nabil Hathout. 2020. ENGLAWI: From human- to machine-readable Wiktionary. In *Proceedings of the twelfth international conference on language resources and evaluation (LREC 2020)*, 3009–3019. Marseille.

Sajous, Franck & Nabil Hathout. 2015. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, 405–426. Herstmonceux, England.

Références

Stroppa, Nicolas & François Yvon. 2005. An analogical learner for morphological analysis. In *Proceedings of the 9th conference on computational natural language learning (conll-2005)*, 120–127. Ann Arbor, MI: ACL.

Štekauer, Pavol. 2014. Derivational paradigms. In Rochelle Lieber & Pavol Štekauer (eds.), *The Oxford handbook of derivational morphology*, 354–369. Oxford: Oxford, Oxford University Press.