LES 20 ANS DE L'ATILF

Identifier les paradigmes morphologiques dérivationnels à partir des dictionnaires électroniques

Nabil Hathout

CLLE, CNRS & Université de Toulouse

9 avril 2021

PROJET FINANCÉ PAR L'ANR
ANR
PROJECT FUNDED BY THE ANR

Partie 1

Les connaissances morphologiques dans les dictionnaires

Peu de ressources morphologiques dérivationnelles du français

Constat

- Dal et al. (1999) constatent qu'il n'existait pas pour le français en 1999 de ressources morphologiques dérivationnelles à large couverture similaires à CELEX (Baayen et al., 1995).
- Un peu plus de 20 ans plus tard, le constat reste globalement vrai. Il n'existe toujours pas d'équivalent de CELEX pour le français.

Les données « primaires » existent

Il existe plusieurs grands dictionnaires électroniques du français.

- Le TLFi (Dendien, 1996; Dendien & Pierrel, 2003) est l'un des premiers (2001).
- Sajous et al. (2014); Sajous & Hathout (2015) ont créé GLAWI en analysant et en balisant en XML les articles du Wiktionnaire.

Inventaires des connaissances morphologiques présentes dans les dictionnaires

Les connaissances morphologiques sont présentes dans :

- les articles consacrés à des affixes et des formants ;
- les entrées secondaires (rubriques Remarques et Dérivés du TLF) ;
- les mots vedettes et les catégories ;
- les définitions ;
- les sections morphologiques (GLAWI) ;
- les rubriques étymologiques.

Cette présentation a bénéficié des discussions avec l'équipe qui a créé le TLFi : Pascale Bernard, Jacques Dendien, Françoise Henry, Christiane Jadelot ainsi qu'avec Jacques Maucourt, Marc Papin parmi beaucoup d'autres.

Entrées consacrées aux affixes et aux éléments formants

Dans le TLF

- *cochonnerie*, *diablerie*, *espièglerie*, *gaminerie* sont des entrées secondaires dans l'article -ERIE, suff.
  ⇒ nom féminin ; suffixation en *-erie* ; la base n'est pas donnée.
- *préformation*, *préretraite*, *précensure*, *prédécoupage* sont des entrées secondaires dans l'article PRÉ-, préf.
  ⇒ nom ; préfixation en *pré-*.
- *infracellulaire*, *infrasensoriel*, *infraorbitaire* sont des entrées secondaires dans l'article INFRA-, élém. formant
  ⇒ composé savant ; adjectif ; *infra* est un des composants.

La distinction préfixe / élément formant a changé dans le TLF à partir du volume 9.

Entrées consacrées aux affixes et aux éléments formants

Dans le Wiktionnaire, les articles sur les préfixes, suffixes et composants contiennent des listes de mots construits par le procédé. Par exemple :

- « Mots en français suffixés avec *-able* »
- « Tableau des entrées du Wiktionnaire se terminant par *able* » (anagrimes).

Ces listes ne sont pas incluses dans GLAWI.

Entrées secondaires

Dans le TLF

Les articles contiennent des rubriques REM (Remarques) et DÉR (Dérivés) qui décrivent des lexèmes apparentés et dérivés de l'entrée principale.

- *invasif* se trouve dans une rubrique REM de l'article INVASION ;
- *vanteur* se trouve dans une rubrique REM de l'article VANTER ;
- *vanterie* se trouve dans une rubrique DÉR de l'article VANTER.

Ces informations sont identifiées explicitement dans le TLFi.

Les entrées secondaires sont spécifiques au TLF.

GLAWI contient des sections morphologiques qui fournissent des mots apparentés, des dérivés et des composés.

Mot vedette et catégorie

Dans le TLF

- Les titres des entrées contiennent la forme de citation du lexème, avec une indication sur les variantes, le féminin, le pluriel et les catégories grammaticales.
  - BONJOURIER, IÈRE, BONJOURIEN, IENNE, subst.
  - BOURGEOIS, OISE, subst. et adj.
  - BRACHIAL, ALE, AUX, adj.
- La forme de citation sert à calculer les relations de formes.
- L'information sur le féminin et le pluriel n'est pas explicite.
- Dans les entrées qui comportent plusieurs catégories, on ne sait pas à laquelle correspondent les descriptions lexicographiques données dans l'article.

GLAWI contient une page pour chaque forme fléchie ;
dans cette page, on trouve une section par catégorie grammaticale.

Définitions

Définitions morphologiques

Les dérivés morphologiques sont généralement définis relativement à un autre membre de leur famille morphologique (Martin, 1983).

- APPROVISIONNEMENT = action d'**approvisionner**.
- CHIRURGICAL = relatif à la **chirurgie**.
- GÉNÉROSITÉ = qualité d'une personne **généreuse**, d'un cœur **généreux**.

- Les définitions des dérivés ne sont pas toutes morphologiques :
  - FACILITÉ = caractère de ce qui se fait sans effort, sans peine.
- Les définitions morphologiques ne sont pas identifiées dans les dictionnaires.
- Le mot de la famille morphologique qui sert à définir le dérivé n'est pas identifié dans les définitions, sauf dans les entrées secondaires.

Définitions

- Les définitions sont correctement identifiées dans les dictionnaires électroniques.
- Elles sont systématiquement présentes dans les articles.
- Les définitions morphologiques décrivent des relations dérivationnelles sémantiquement motivées.

$\Rightarrow$ Les définitions sont une source très fiable de connaissances morphologiques.

Rubriques étymologiques

Les rubriques étymologiques contiennent généralement une description explicite des procédés morphologiques qui ont servi à construire les lexèmes morphologiquement complexes.

houppette

Étymol. et Hist. : 1399 « petite houppe » (doc. ds GDF. Compl.). Dér. de *houppe**; suff. dimin. *-ette**.

Rubriques étymologiques

humoristique

Étymol. et Hist. I. 1801 (CRAMER ds S. MERCIER, Néol., t. 1, p. 332 : les expressions les plus délicates, **Humoristiques** plaisantes du sentiment et de l'imagination); 1869 *humouristique* (STE-BEUVE, loc. cit.). **II.** 1847 « des humeurs » *mécanisme humoristique de l'homme* (BALZAC, loc. cit.). Dér. de *humoriste**; suff. *-ique**. I en rapport avec l'angl. *humoristic* (attesté seulement dep. 1818 ds NED) et *humour**. II d'apr. *humeur**.

Les rubriques étymologiques ne sont pas analysées dans les dictionnaires électroniques car :

- elles n'ont pas une structure stable
- leur contenu est très variable.

Elles contiennent à la fois des informations historiques et morphologiques

Rubriques étymologiques

Les relations morphologiques des emprunts ne sont pas décrites, même lorsqu'elles s'établissent clairement en synchronie.

formation

Étymol. et Hist. 1. *Ca* 1170 *la formation deu monde* (BENOIT, *Chronique des Ducs de Normandie*, éd. C. Fahlin, 47); 1774 géol. (BUFFON, *Hist. nat.*, t. 1, p. 92); 2. 1789 *formation des sociétés politiques* (SIEYÈS, *Tiers état*, p. 65); 1790 milit. (MARAT, *Pamphlets*, p. 82 : la **formation** des camps au Champ de Mars); 3. 1898 « éducation d'un être humain » *ma formation littéraire* (BARRÈS, *Cahiers*, t. 1, p. 25). Empr. au lat. *formatio* « forme, confection » cf. la forme avec traitement semi-pop. du suff. *formaison* (XIIe s. ds T.-L.) en partic. en gramm. « formation des temps » (XIIIe s., *ibid.*).

⇒ Les rubriques étymologiques ne sont pas exploitables.

Bilan

- Les dictionnaires contiennent l'essentiel de l'information nécessaire à la description de la morphologie.
- L'information est présente dans un grand nombre d'objets lexicographiques.
- Elle se présente sous une forme spécifique dans chaque type d'objet.
- Elle est explicite dans les rubriques étymologiques, mais ces rubriques sont totalement inexploitables.
- Elle est implicite et massivement présente dans les définitions (Hathout, 2009b, 2011a; Hathout et al., 2014).
  Les définitions sont parmi les objets les mieux analysés dans les dictionnaires électroniques.

Remarques

- Les dictionnaires peuvent être utilisés comme des corpus pour construire des représentations vectorielles des définitions (Hill et al., 2016; Bosc & Vincent, 2018).
  Dans la même lignée, les travaux réalisés à l'ATILF par Mickus et al. (2019) ont pour finalité la génération automatique de définitions.
- L'utilisation des dictionnaires dépend de leur disponibilité sous forme électronique, de la qualité de leur analyse, mais aussi des licences sous lesquelles ils sont diffusés.
  Les recherches en TAL sont aujourd'hui de plus en plus limités aux seuls jeux de données disponibles publiquement sous licences libres.

Partie 2

Acquisition de connaissances morphologiques à partir de dictionnaires électroniques

En collaboration avec **Basilio Calderone** (CLLE)
Franck Sajous (CLLE)
Fiammetta Namer (ATILF)

Paradigmes flexionnels

- La morphologie flexionnelle du français est paradigmatique.
- Les formes fléchies des verbes sont décrites au moyen de tables de conjugaison (Beschrelle).

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	...

Paradigmes flexionnels

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

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

Morphologie dérivationnelle paradigmatique

La description paradigmatique de la morphologie dérivationnelle permet de rendre de compte simplement de nombreux phénomènes non canoniques (Bochner, 1993; Bauer, 1997; Štekauer, 2014; Antoniova & Štekauer, 2015; Blevins, 2016; Hathout & Namer, 2018a,b; Bonami & Strnadová, 2019; Hathout & Namer, 2019; Namer & Hathout, 2020)

La transposition à la morphologie dérivationnelle des tables flexionnelles sont des tables dérivationnelles :

Paradigme dérivationnel (partiel)

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

Morphologie dérivationnelle paradigmatique

- Les lignes de la table décrivent des familles morphologiques
  = des ensembles de lexèmes morphologiquement apparentés.
- Les colonnes de la table décrivent des séries morphologiques
  = des ensembles de lexèmes dont les contrastes de forme et de sens avec les autres membres de leurs familles morphologiques sont identiques (Roché, 2009; Hathout, 2009c, 2011b).

Glawinette (Hathout et al., 2020)

Glawinette est un lexique morphologique extrait du dictionnaire électronique GLAWI (Sajous & Hathout, 2015; Hathout & Sajous, 2016).

Glawinette

- 97 293 lexèmes
- 47 717 couples de mots morphologiquement apparentés
- 15 904 familles dérivationnelles
- 5400 séries de couples morphologiques

Glawinette (Hathout et al., 2020)

Glawinette est un lexique relationnel.

- 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 (Hathout et al., 2020)

- Les séries morphologiques sont des ensembles de couples morphologiques qui présentent les mêmes contrastes de forme.

Série de couples morphologiques

^(.+)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 (Hathout et al., 2020)

Démonette

Glawinette servira 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).

Glawitina

Un lexique identique a été construit pour l'italien en utilisant le dictionnaire électronique italien GLAWIT (Calderone et al., 2016).

Glawitina

- 14 804 lexèmes
- 19 786 couples de mots morphologiquement apparentés
- 5317 familles dérivationnelles
- 1178 séries de couples morphologiques

Un lexique dérivationnel de l'anglais est en cours de finalisation. EnGlawinette est créé à partir du dictionnaire électronique anglais ENGLAWI (Sajous et al., 2020).

Famille morphologique de *serrer* dans Glawinette

desserrage=N:desserrer=V desserre=N:desserrer=V
desserrement=N:desserrer=V desserrer=V:desserrage=N
desserrer=V:desserre=N desserrer=V:desserrement=N
desserrer=V:desserroir=N **desserrer=V:indesserrable=A**
desserrer=V:redesserrer=V **desserrer=V:serrer=V** desserroir=N:desserrer=V
enserrement=N:enserrer=V enserrer=V:enserrement=N
enserrer=V:renserrer=V enserrer=V:réenserrer=V enserrer=V:serre=N
indesserrable=A:desserrer=V redesserrer=V:desserrer=V
renserrer=V:enserrer=V reresserrer=V:resserrer=V resserrage=N:resserrer=V
resserrement=N:resserrer=V resserrer=V:reresserrer=V
resserrer=V:resserrage=N resserrer=V:resserrement=N
resserrer=V:resserreur=N 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 couples morphologiques dans Glawinette

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

Caractérisation des couples morphologiques dans Glawinette

- Les relations morphologiques sont symétrisées.
- Les relations morphologiques sont à la fois directes et indirectes.
- Les schémas peuvent être utilisés génératívement pour prédire des dérivés.
- Les schémas mettent en jeu des exposants aussi proches que possible de ceux que les linguistes utilisent pour décrire les procédés morphologiques.

Exposants « naturels »

formalisme:formaliser	`^(.+)isme$=N`	`^(.+)iser$=V`
formatif:formation	`^(.+)atif$=A`	`^(.+)ation$=N`
formellement:formel	`^(.+)ellement$=R`	`^(.+)el$=A`
réforme:réformette	`^(.+)e$=N`	`^(.+)ette$=N`
réformiste:réformisme	`^(.+)iste$=A`	`^(.+)isme$=N`

Les étapes de la construction de Glawinette

1. Les couples extraits des sections morphologiques ne sont pas totalement fiables.
2. Les définitions fournissent des couples sémantiques bruités.
3. L'analogie permettent d'identifier des régularités formelles et les couples de lexèmes qui sont formellement motivées.
4. Les régularités formelles permettent de constituer des séries morphologiques.
5. L'analogie permet d'identifier les régularités formelles des couples de lexèmes qui composent des séries morphologiques et d'éliminer les couples les moins réguliers.
6. Les familles morphologiques sont des graphes connexes de couples de lexèmes.
7. Les exposants les plus « naturels » maximisent le nombre de connexions des lexèmes qu'ils décrivent dans l'ensemble du lexique.

1. Extraction des couples présents dans les sections morphologiques

GLAWI contient 32 249 sections morphologiques à partir desquelles nous avons extraits 125 002 couples morphologiques entre des lexèmes

- dont le lemme est typographiquement simples (sans espace, tiret, apostrophe, etc.),
- dont le lemme ne contient que des lettres minuscules
- qui appartiennent aux catégories nom, verbe, adjectif, adverbe.

Couples morphologiques

motorisation:motoriser ; anticolonialisme:colonialisme ; colonialisme:colon ; stéréocomparateur:comparateur ; distorsion:distordre

Couples non morphologiques

âme:éprouver ; électronique:sélecteur ; femme:toilette

2. Extraction des couples présentes dans les définitions

GLAWI contient 355 676 définitions.

Définitions morphologiques

- **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**

2. Extraction des couples présentes dans les définitions

Définitions non morphologiques

- développement = nombre de mètres qu'une bicyclette parcourt sur un coup de pédale
- productif = qui est effectivement utilisé en pratique pour créer des nouveaux mots, par exemple en parlant d'un préfixe ou d'un suffixe, d'une figure de style...

.

- On ne sait pas identifier a priori les définitions morphologiques (Hathout, 2008, 2009a, 2014)
- Dans les définitions morphologiques, on ne sait pas quels sont les lexèmes apparentés à l'entrée (*definiendum*).
- Les définitions de GLAWI sont analysées en dépendance et lemmatisées.

2. Extraction des couples présentes dans les définitions

Méthode

1. On forme **tous** les couples de lexèmes composés de l'entrée et d'un mot de sa définition (*definiens*) tels que
   - les deux lexèmes ont des lemmes typographiquement simples,
   - les deux lexèmes ne contiennent que des lettres minuscules
   - les deux lexèmes appartiennent aux catégories nom, verbe, adjectif, adverbe.
2. On ne conserve parmi ces les couples que ceux dont les contrastes de forme sont suffisamment réguliers
   - qui apparaissent au moins **5 fois** dans les couples extraits des définitions et des sections morphologiques.

3. Identifier les contrastes de forme réguliers au moyen de l'analogie formelle

- Les contrastes de forme entre les couples de lexèmes morphologiquement apparentés tendent à être réguliers (exception : *naïf*:*naïveté*).
- Les couples de mots qui sont dans les mêmes relations formelles forment des analogies.
- Une analogie est un quadruplet $A{:}B{=}C{:}D$ telle que A est à B ce que C est à D.

3. Identifier les contrastes de forme réguliers au moyen de l'analogie formelle

développement*:*développer*=*classement*:*classer

Il existe une factorisation telle que dans chaque colonne, la différence entre les deux premières lignes est identique à celle qu'il y a entre les deux dernières lignes (Lepage, 1998, 2004a,b; Stroppa & Yvon, 2005; Langlais & Yvon, 2008):

d	é	v	e	l	o	p	p	e	m	e	n	t
d	é	v	e	l	o	p	p	e	r	ϵ	ϵ	ϵ
ϵ	ϵ	ϵ	c	l	a	s	s	e	m	e	n	t
ϵ	ϵ	ϵ	c	l	a	s	s	e	r	ϵ	ϵ	ϵ
$=$	$=$	$=$	$=$	$=$	$=$	$=$	$=$	$=$	m/r	e/ϵ	n/ϵ	t/ϵ

3. Identifier les contrastes de forme réguliers au moyen de l'analogie formelle

- Pour identifier les quadruplets analogiques, nous utilisons la méthode proposée par Lepage (1998, 2004b).
  - si $A{:}B{=}C{:}D$ alors la distance de Levenshtein (1966) entre A et B et entre C et D sont égales.
  - 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 une signature

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

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.

- Si $\sigma(A, B) = \sigma(C, D)$ alors $A{:}B{=}C{:}D$ à de très rares exceptions près : *stressed*:*desserts*≠*reward*:*drawer* (Lepage, 2004b)

3. Identifier les contrastes de forme réguliers au moyen de l'analogie formelle

- Les contrastes dans les couples non morphologiques issus des définitions comme *développement*:*action*, *perforation*:*action*, *vernissage*:*action*, etc. sont tous différents.
  - Ils ne sont pas réguliers.
  - Ils ne forment pas d'analogies avec d'autres couples de lexèmes (sauf accident).
- Les contrastes dans les couples *développement*:*développer*, *classement*:*classer*, *chargement*:*charger*, *arasement*:*araser*, etc. sont formellement identiques.
  - Ils sont réguliers.
  - Ils forment une série de 1247 couples.

$\Rightarrow$ On ne conserve que les couples issus des définitions et des sections morphologiques qui font partie d'une série d'au moins 5 couples.

- 170 370 couples sont conservés sur les 2 353 959 initialement formés.

4. Trouver les patrons des séries

- Les analogies sont basées sur un alignement local de 4 lemmes.
- Les couples morphologiques d'une série sont formellement homogènes (Fam & Lepage, 2018).
- Les régularités formelles des couples morphologiques définissent les exposants des lexèmes qu'ils réunissent.
- Les régularités formelles des couples morphologiques peuvent être décrites au moyen de patrons.

4. Trouver les patrons des séries

Régularités formelles dans les séries de couples morphologiques

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

- Les lemmes de la 1re colonne finissent tous en *-eur* : `eur$`
- Les lemmes de la 2e colonne finissent tous en *-age* : `age$`
- Dans chaque couple, les parties de deux lemmes suffixés en *-eur* et en *-age* sont identiques : `^(.+)`

4. Trouver les patrons des séries

Séparer les séries fusionnées

- Les couples d'une série formelle qui instancient des patrons différents appartiennent à des séries dérivationnelles différentes.
- Les couples des séries *-eur*/*-age* et *-ure*/*-age* ont les mêmes signatures

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

⇒ *allumeur*:*allumage* et *doublure*:*doublage* ont la même signature, mais *allumeur*:*allumage*≠*doublure*:*doublage*.

4. Trouver les patrons des séries

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 mettant en correspondance les patrons de mots dont les parties variables (`.+`) sont instanciées par la même chaîne de caractères dans les couples de mots de la série.

4. Trouver les patrons des séries

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

Patron *-eur*	Cov
^(.+)$	1.0
^(.+)u$	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

4. Trouver les patrons des séries

Trouver les patrons d'une série formelle

- On compare tous les couples de la série deux-à-deux
- On calcule pour chaque paire 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`)
  - Le patron est étendu en ajoutant une partie de longueur identique des sous-chaînes initiales ou finales des radicaux
  - Ce découpage permet de construire l'ensemble des patrons qui décrivent (w_1, w_2, w_3, w_4)
- On conserve les patrons qui :
  - ne contiennent qu'un seul radical `(.+)`
  - décrivent au moins 5 couples
  - couvrent au moins 10% de la série formelle.

5. Trouver les exposants les plus naturels

- Les couples retenus sont souvent décrits par plusieurs patrons.
- Nous proposons une méthode qui permet de choisir le patron dont les exposants sont les plus « naturels »

		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$:^(.+)sme$`	←

5. Trouver les exposants les plus naturels

- 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 « naturalité » 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.

5. Trouver les exposants les plus naturels

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 lexème $Z \in \{X_1, X_2, ...\} \cup \{Y_1, Y_2, ...\}$, 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.

Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers. 1995. The CELEX lexical database (release 2). CD-ROM. Linguistic Data Consortium, Philadelphia, PA.

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.

Références

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

Bosc, Tom & Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 1522–1532.

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.

Dal, Georgette, Nabil Hathout & Fiammetta Namer. 1999. Construire un lexique dérivationnel : Théorie et réalisation. In Pascal Amsili (ed.), *Actes de la 6^{e} conférence sur le traitement automatique des langues naturelles (taln-99)*, 115–124. Cargèse: ATALA.

Dendien, Jacques. 1996. Le projet d'informatisation du tlf. In David Piotrowski (ed.), *Lexicographie et informatique. autour de l'informatisation du Trésor de la Langue Française*, 25–34. Paris: Didier Érudition.

Références

Dendien, Jacques & Jean-Marie Pierrel. 2003. Le trésor de la langue française informatisé: un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement automatique des langues* 44(2). 11–37.

Fam, Rashel & Yves Lepage. 2018. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 1060–1066. Miyazaki, Japan.

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.

Hathout, Nabil. 2009a. Acquisition morphologique à partir d'un dictionnaire informatisé. In *Actes de la 16^e 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.

Références

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.

Hathout, Nabil. 2014. Phonotactics in morphological similarity metrics. *Language Sciences* 46. 71–83.

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

Références

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.

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`.

Références

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. 2014. Acquisition and enrichment of morphological and morphosemantic knowledge from the french wiktionary. In *Proceedings of the coling workshop on lexical and grammatical resources for language processing*, 65–74. Dublin, Ireland.

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.

Hill, Felix, Kyunghyun Cho, Anna Korhonen & Yoshua Bengio. 2016. Learning to Understand Phrases by Embedding the Dictionary. *Transactions of the Association for Computational Linguistics* 4. 17–30.

Références

Langlais, Philippe & François Yvon. 2008. Scaling up analogical learning. In *Proceedings of the 22nd international conference on computational linguistics (coling 2008)*, 51–54. Manchester.

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.

Levenshtein, Vladimir Iosifovich. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady* 10(8). 707–710.

Martin, Robert. 1983. *Pour une logique du sens* Linguistique nouvelle. Paris: Presses universitaires de France.

Références

Mickus, Timothee, Denis Paperno & Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the first NLPL workshop on deep learning for natural language processing*, 1–11. Turku, Finland.

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-2003)taln*, 233–243. Toulouse.

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.

Références

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.

Sajous, Franck, Nabil Hathout & Basilio Calderone. 2014. Ne jetons pas le Wiktionnaire avec l'oripeau du web ! Études et réalisations fondées sur le dictionnaire collaboratif. In *Actes du 4e Congrès Mondial de Linguistique Française (cmlf 2014)*, 663–680. Berlin. Germany.

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.