Prédire les relations morphologiques à partir de la forme et du sens

Basilio Calderone, Nabil Hathout

CLLE, CNRS & Université Toulouse Jean Jaurès

Thématiques actuelles de la recherche en TAL
21 novembre 2022

Prédire les lexèmes de la famille dérivationnelle

Caractérisation formelle des relations dérivationnelles

Le lexique Glawinette

Questions

Méthodes et données

Expériences

Conclusion

Paradigm Cell Filling Problem

Objectif : apporter une contribution à la résolution du **Paradigm Cell Filling Problem** (PCFP) pour la morphologie dérivationnelle.

PCFP repose sur l'hypothèse que la morphologie flexionnelle organise les formes fléchies des lexèmes en **paradigmes**.

Les paradigmes flexionnels s'appellent des **classes flexionnelles**.

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.

Paradigm Cell Filling Problem

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.

Paradigm Cell Filling Problem

Les locuteurs ne sont pas exposés à toutes les formes fléchies de tous les lexèmes présents dans le lexique. Néanmoins, ils sont capables de produire la famille flexionnelle de n'importe quel lexème à partir d'une seule des formes de ce lexème.

Un locuteur qui entend :

Les enfants datoperont plus tard

... saura produire les formes de l'imparfait du verbe :

je datopais	nous datopions
tu datopais	vous datopiez
il/elle datopait	ils/elles datopaient

... c.à.d remplir les cases correspondantes dans sa famille flexionnelle.

Paradigm Cell Filling Problem

PCFP est une tâche qui consiste à prédire l'ensemble des formes fléchies d'un lexèmes sur la base de l'une d'entre elles.

Paradigm Cell Filling Problem

> *Given exposure to an inflected wordform of a novel lexeme, what licenses reliable inferences about the other wordforms in its inflectional family?* (Ackerman et al., 2009)

Malouf (2017, 2016) a proposé une méthode qui permet de résoudre PCFP au moyen de réseaux récurrents LSTM. Les performances du modèle dépassent celles des locuteurs humains : **99.94% d'exactitude** pour le français.

Transposer PCFP à la dérivation

Hypothèse 1. la morphologie dérivationnelle organisent les lexèmes en familles dérivationnelles et que ces familles s'organisent à leurs tours en paradigmes (Bochner, 1993; Bauer, 1997; Štekauer, 2014; Antoniova & Štekauer, 2015; Blevins, 2016; Hathout & Namer, 2018a,b; Bonami & Strnadová, 2019; Hathout & Namer, 2019; Namer & Hathout, 2020; Hathout & Namer, 2022).

Paradigme dérivationnel

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

Transposer PCFP à la dérivation

- Les lignes décrivent des **familles morphologiques** = ensembles de lexèmes morphologiquement apparentés (Roché, 2009; Hathout, 2009, 2011; Roché, 2017; Fradin, 2020, 2021).
- Les colonnes décrivent des **séries morphologiques de lexèmes** = ensembles lexèmes qui présentent les mêmes contrastes de forme et de sens avec les autres membres de leurs familles morphologiques.
- Les couples de colonnes décrivent des **relations morphologiques** = ensemble de couples de lexèmes qui présentent les mêmes contrastes de forme et de sens.

Hypothèse 2. Les paradigmes dérivationnels sont définis sur la base des relations sémantiques et catégorielles entre les lexèmes des familles.

Lexèmes		Relation
laver	laveur	‘@’$_V$:‘celui qui @’$_{Nm}$
laver	lavable	‘@’$_V$:‘que l’on peut @’$_A$
laveur	lavable	‘celui qui @’$_{Nm}$:‘que l’on peut @’$_A$

Transposer PCFP à la dérivation

La transposition de PCFP à la dérivation :

> *Étant donné un lexème d'une nouvelle famille dérivationnelle, prédire les autres lexèmes de cette famille ?*

Pour adapter entièrement PCFP à la dérivation, il faut savoir :

- à quel paradigme appartient la famille dérivationnelle de chaque lexème du lexique ?
- quels sont les lexèmes qui composent cette famille (qu'elles sont les cellules du paradigme) ?
- quelles sont les relations sémantiques et catégorielles qui s'établissent entre ces lexèmes ?
- comment ces relations sont-elles caractérisées ?

Transposer PCFP à la dérivation

Cette transposition a été proposée par Bonami & Strnadová (2019); Boyé & Schalchli (2019) parmi d'autres.

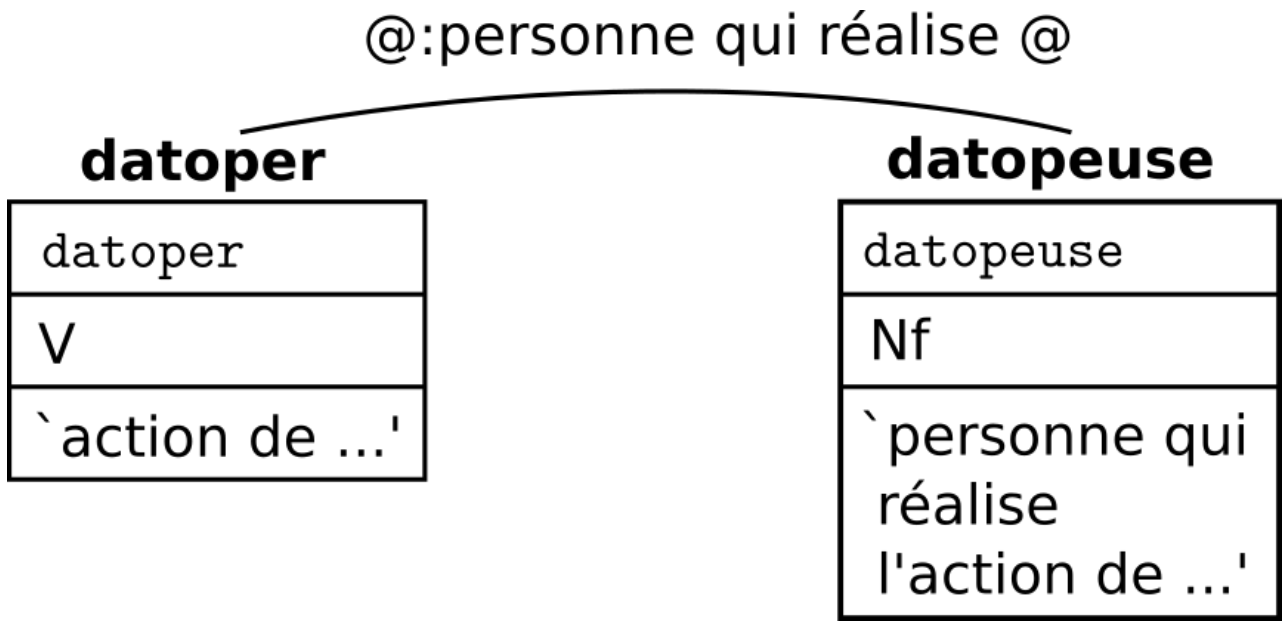
Cotterell et al. (2017) ont proposé une solution pour une version du PCFP dérivationnel basée sur des réseaux de neurones seq2seq. L'étude a porté sur l'anglais et sur des familles dérivationnelles composées d'un verbe d'action et de 4 dérivés identifiés par les catégories sémantiques : AGENT, PATIENT, RESULT, POTIENTIAL.

Transposer PCFP à la dérivation

WORK IN PROGRESS

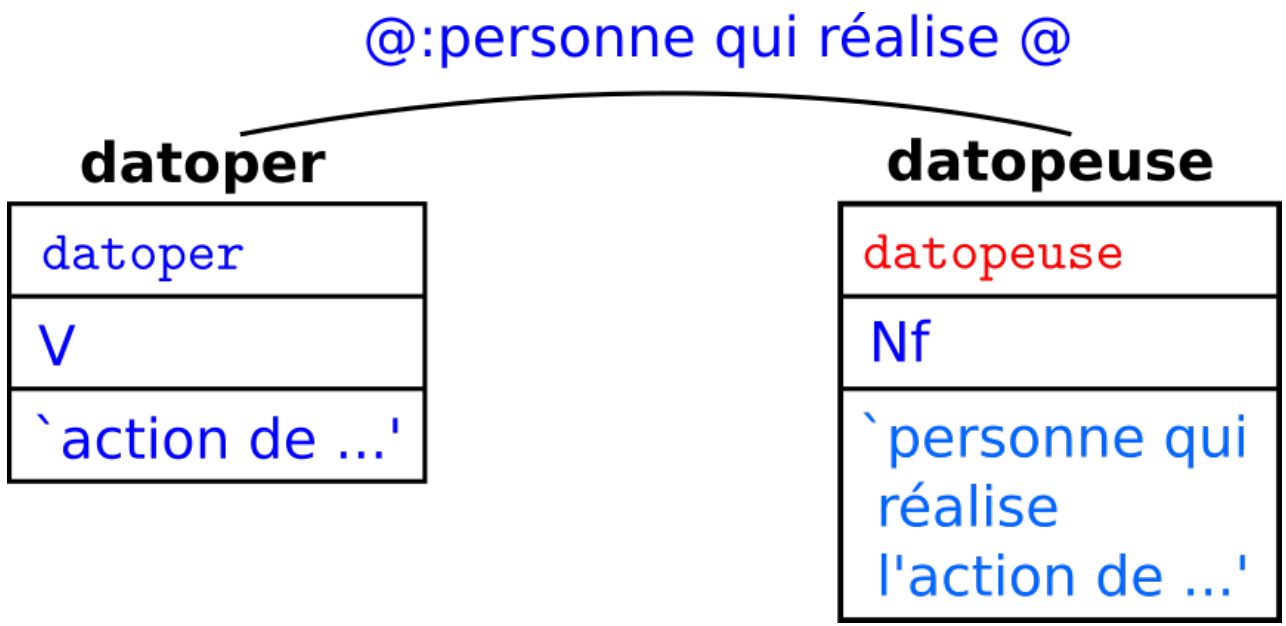
Nous nous proposons de réaliser une tâche plus simple que PCFP :

> *Soit L1 lexème et R une relation sémantique pouvant relier L1 à un autre lexème L2 de sa famille dérivationnelle, prédire le lexème L2*



FAP

PCFP en dérivation : Si l'on suppose que la relation sémantique permet de définir le sens du lexème cible, seule la forme de ce dernier est à calculer.



FAP

Nous proposons de caractériser les relations morphologiques au moyen de patrons d'alternance (*fine-graned alternance patterns*, FAP) (Hathout et al., 2020; Calderone et al., 2021).

Un FAP est un couple de patrons décrits sous forme d'expressions régulières. Dans ces expressions régulières les sous-expressions `(.+)` correspondent aux mêmes séquences.

Lexèmes		FAP1	FAP2	`(.+)`
laver	laveur	`^(.+)er$`	`^(.+)eur$`	lav
préfiltrer	filtrage	`^pré(.+)er$`	`^(.+)age$`	filtr
réalisme	réaliste	`^(.+)isme$`	`^(.+)iste$`	réal

La connaissance de la forme de l'un des lexèmes et du FAP qui le relie au second lexème permet de prédire la forme de ce dernier.

FAP

Les FAP ont généralement une bonne validité linguistique. Ils sont construits à partir des signatures analogiques des séries de lexèmes qui composent les relations morphologiques (séries de couples de lexèmes).

Ils sont sélectionnés par une méthode similaire à la minimisation de la longueur de description (MDL) (Goldsmith, 2001, 2006).

Les FAP sont des descriptions qui s'avèrent adaptés à l'acquisition des régularités morphologiques au moyen de réseaux de neurones (Calderone et al., 2021).

Les FAP constituent un bon compromis entre généralité et spécificité / informativité.

Version simplifier du PCFP dérivationnel

Prédire le patron (FAP2) du lexème cible.

Glawinette

Glawinette est un lexique dérivationnel du français construit à partir du dictionnaire électronique GLAWI (Sajous & Hathout, 2015; Hathout & Sajous, 2016; Hathout et al., 2020).

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

Glawinette

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

`http://redac.univ-tlse2.fr/lexiques/glawinette.html`

Définitions morphologiques

Glawinette a été construit à partir des sections morphologiques et des définitions morphologiques de GLAWI.

- 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.

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**

Jeu de données

Notre jeu de données est construit à partir des entrées de Glawinette :

- issues de définitions morphologiques ;
- dont les définitions contiennent entre 2 et 15 mots (sans ponctuation).

Le jeu de données contient **44 550 entrées**.

Définition	Lemme1	Lemme2	C1	C2	FAP1	FAP2	Rad
['action', 'de', 'écraser']	écraser	écrasage	V	N	+er	+age	écras
['relatif', 'le', 'production', 'ou', 'le', 'commerce', 'de', 'le', 'drap']	drap	drapier	N	A	+	+ier	drap
['spécialiste', 'de', 'le', 'logique', 'en_tant_que', 'discipline']	logique	logicien	N	N	+ique	+icien	log

Dans les FAP, les sous-expressions `(.+)` sont simplifiées en `+`.

Question 1: Difficulté de la tâche

Dans quelle mesure peut-on prédire le FAP du lexème cible à partir des différentes informations (formelles, catégorielles et sémantiques) qui décrivent le lexème source et la cellule cible (relation sémantique, catégorie) ?

$\text{Lemme}_{\text{source}}$	$\text{Cat}_{\text{source}}$	$\text{Sens}_{\text{source}}$	FAP
livrer	V	livrer	+er

↓

$\text{Lemme}_{\text{cible}}$	$\text{Cat}_{\text{cible}}$	$\text{Sens}_{\text{cible}}$	FAP
livreur	N	celui, celle qui **livre** des marchandises, des colis, au nom d'une compagnie	+eur

Question 1: Difficulté de la tâche

Pour répondre à la question, nous construisons des modèles qui disposent de quatre types d'information :

1. la forme du lemme source (**livrer**) ;
2. la catégorie du lexème cible (**N**) ;
3. le sens du lexème cible (**celui, celle qui livre des marchandises, des colis, au nom d'une compagnie**)
4. le FAP du lexème cible (**+eur**)

Question 2 : Contribution de la forme du lemme source

Quelle est la contribution de la forme du lemme source à la prédiction de FAP du lexème cible ?
L'omission de la forme du lemme source détériore-t-elle la prédiction du FAP du lexème cible ?

Pour répondre à la question, nous construisons des modèles qui disposent de trois types d'information :

1. ~~la forme du lemme source (**livrer**) ;~~
2. la catégorie du lexème cible (**N**) ;
3. le sens du lexème cible (**celui, celle qui livre des marchandises, des colis, au nom d'une compagnie**)
4. le FAP du lexème cible (**+eur**)

Question 3 : Contribution du sens du lexème source

Le sens du lexème source est contenu dans la description du sens du lexème cible (définition).

En le supprimant de cette description, la définition devient une caractérisation de la relation sémantique entre le lexème source et le lexème cible. Cette relation caractérise la cellule du lexème cible dans le paradigme dérivationnel.

Quelle est la contribution du sens du lexème source à la prédiction du FAP du lexème cible ?
Est-ce que la prédiction du FAP du lexème cible est détériorée si l'on remplace la description du sens du lexème cible par celle de la relation sémantique entre les deux lexèmes ?

Question 3 : Contribution du sens du lexème source

Pour répondre à la question, nous construisons des modèles qui disposent de quatre types d'information :

1. la forme du lemme source (**livrer**) ;
2. la catégorie du lexème cible (**N**) ;
3. la relation sémantique qui caractérise la cellule du lexème cible (**celui, celle qui ~~livre~~ des marchandises, des colis, au nom d'une compagnie**)
4. le FAP du lexème cible (**+eur**)

Traiter des séquences

Les informations que nous utilisons pour la prédiction du FAP sont de nature séquentielle :

- **La représentation orthographique du lemme source** est une chaîne de caractères de longueur variable

$$L \mapsto I \mapsto V \mapsto R \mapsto E \mapsto R$$

- **La définition du lexème cible** est une séquence ordonnée de mots qui chacun apporte un contenu sémantique.

celui $\mapsto$ celle $\mapsto$ qui $\mapsto$ livrer $\mapsto$ de $\mapsto$ le $\mapsto$ marchandise ...

Nous avons développé une architecture neuronale qui utilise des modèles LSTM. Ces modèles sont adaptés au traitement des séquences.

Traiter des séquences

Long Short-Term Memory. Un système **LSTM** (Hochreiter & Schmidhuber, 1997) est un modèle neuronal capable d'**encoder** des données en séquences (textes, vidéos, etc.), de sélectionner les informations les plus importantes qu'elles contiennent et d'écarter les séquences (ou parties de séquences) les moins importantes.

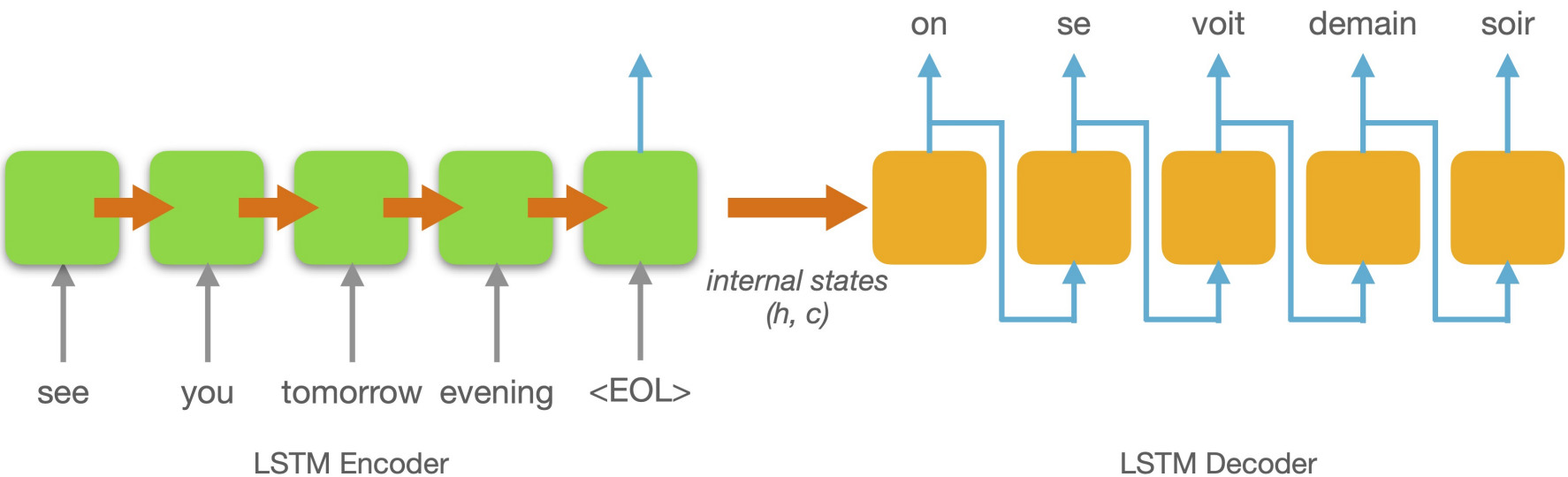
Les séquences d'entrée sont utilisées par un **encodeur** pour construire une représentation interne (matrice).

La représentation interne est utilisée par un **décodeur** pour produire une séquence de sortie.

Exemple de LSTM pour la traduction

1. L'encodeur (carrés verts) lit de gauche à droite, un par un les éléments de la séquence d'entrée.
2. L'encodeur crée une représentation matricielle de la séquence qui correspond à l'état d'activation interne du système lorsque la lecture de la séquence est terminée (grande flèche orange).
3. La matrice est transmise au décodeur (carrés jaunes) qui la décode en une nouvelle séquence.

Séquences de *word embeddings*



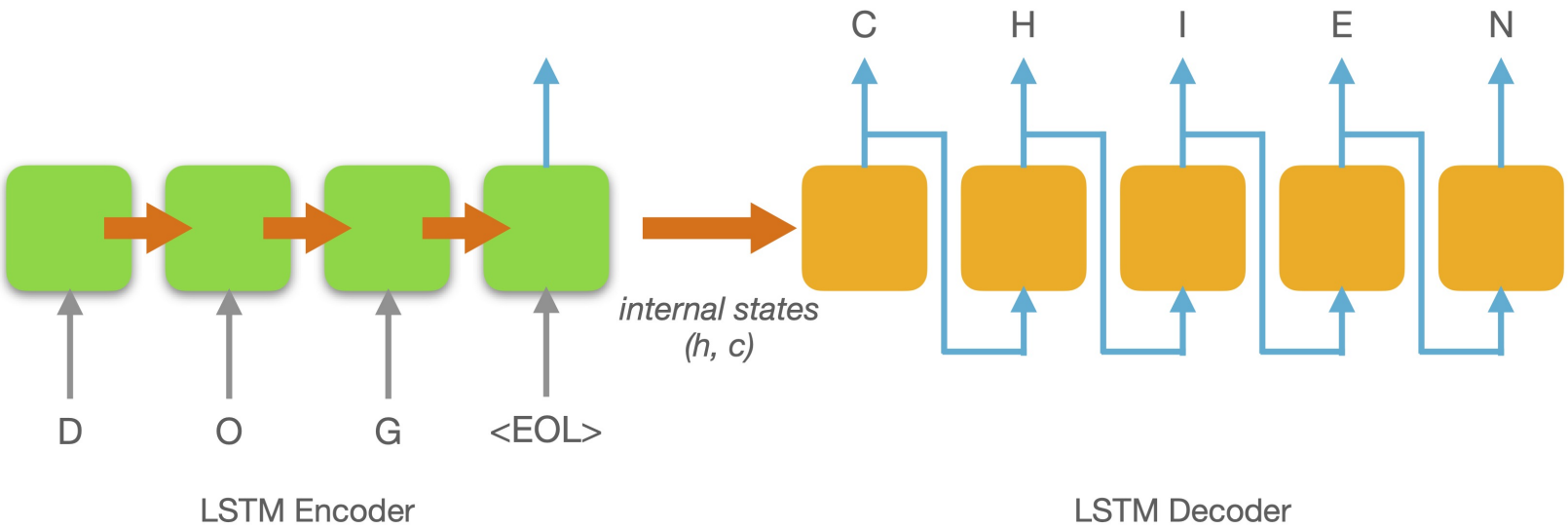
Types de séquences temporelles

L'architecture encodeur/décodeur permet de traiter des séquences de différents types : séquences de pixels, de caractères, de *word embeddings*.

Les séquences en entrée et en sortie peuvent être de natures différentes.

Les séquences peuvent être hétérogènes.

Séquences de caractères

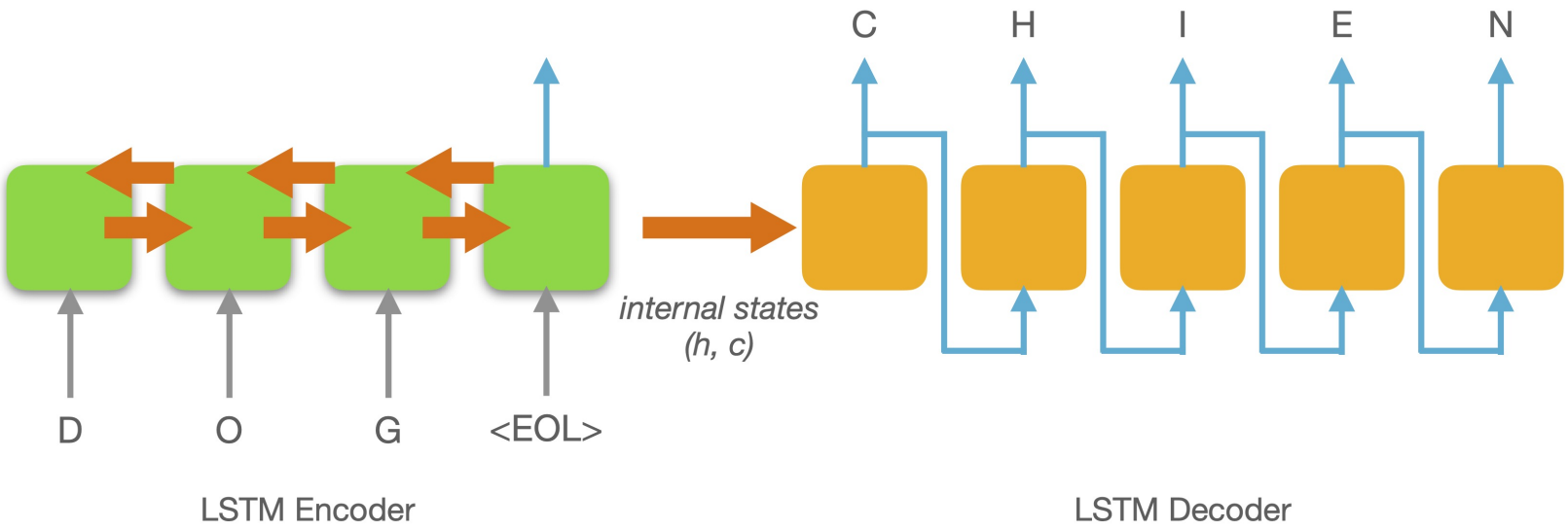


Encodeur bi-directionnel : Bi-LSTM

Il est possible d'améliorer les performances du modèle en utilisant un encodeur bi-directionnel.

Un encodeur bi-directionel crée deux représentations, l'une en lisant la séquence de gauche à droite et l'autre en la lisant de droite à gauche.

Les deux représentations sont ensuite concaténées.



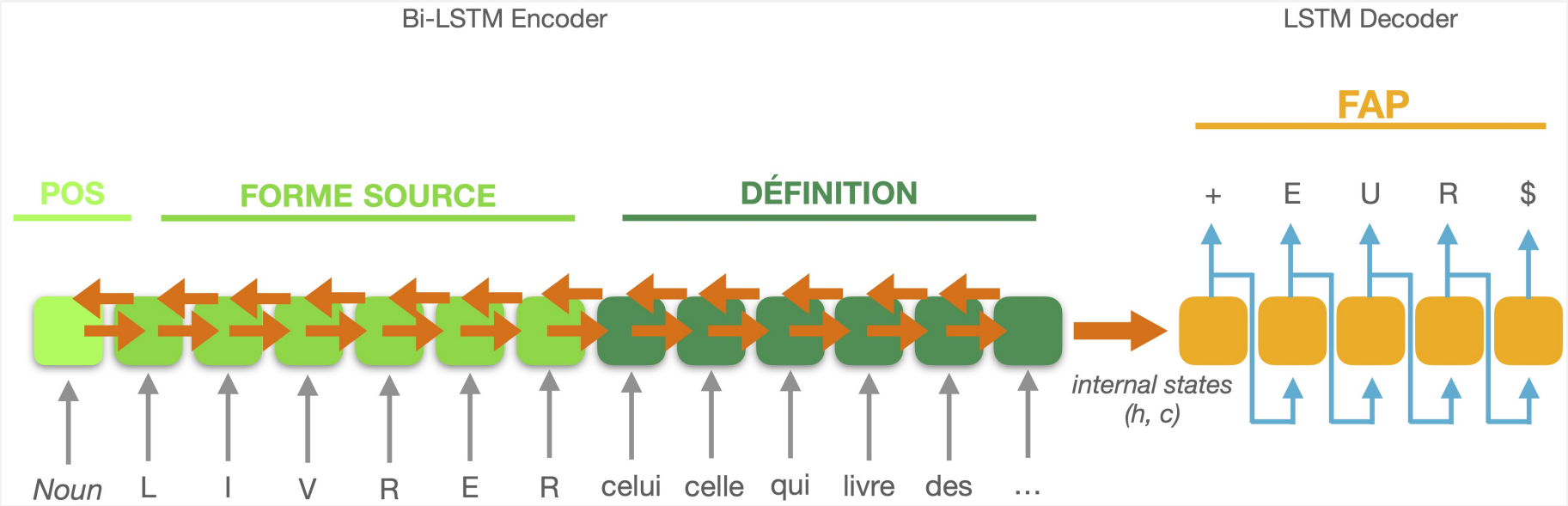
Nos modèles

Modèle « à séquence unique »

On concatène les trois types d'informations (formelle, catégorielle et sémantique) en une seule séquence d'entrée.

Cette séquence est l'entrée d'un bi-encodeur.

Le décodeur produit en sortie le FAP du lexème cible.



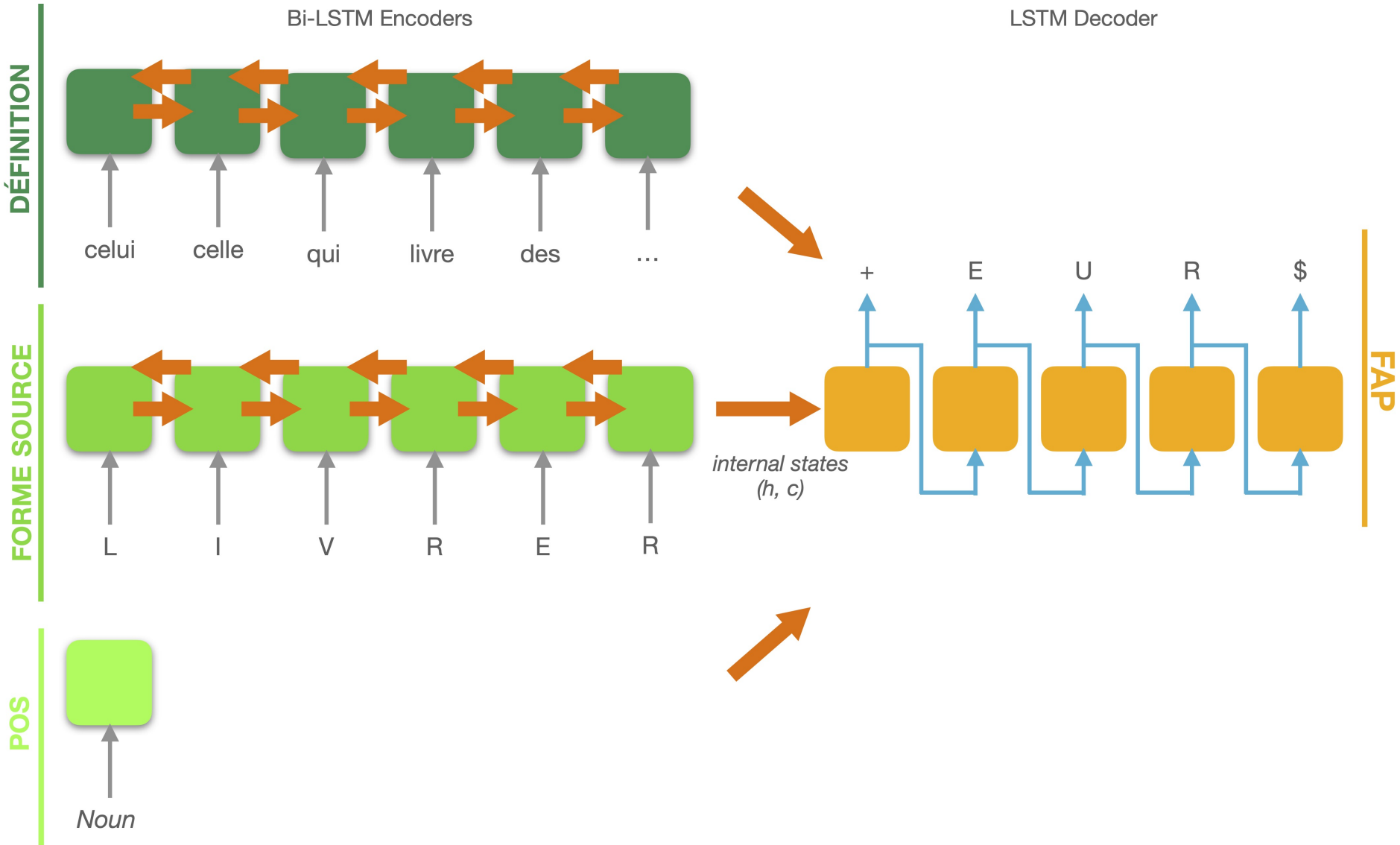
Nos modèles

Modèle « à concatenation »

Chaque séquence est traitée par un bi-encodeur dédié. Les représentations matricielles créés par les trois encodeurs sont concaténées puis transmises au décodeur.

Le décodeur produit en sortie le FAP du lexème cible.

Nos modèles


Bi-LSTM Encoders
LSTM Decoder
DÉFINITION
FORME SOURCE
POS
celui
celle
qui
livre
des
...
L
I
V
R
E
R
Noun
internal states
(h, c)
+
E
U
R
$
FAP

Données et paramètres

Jeu de données

Entrées de Glawinette construites à partir de définitions morphologiques

Nous éliminons les entrées dont les définitions sont trop courtes ou trop longues ($2 \leq \text{len(Définition)} \leq 15$)

44 550 définitions au total

Paramètres

- 37 866 entrées sont utilisées pour l'entraînement **Training set**
- 6 682 entrées sont utilisées pour le test **Test set**
- 150 neurones dans chaque encodeur et chaque décodeur
- 200 époques d'apprentissage

Représentation des catégories et des caractères

La catégorie et les caractères du lemme source sont encodés en **one-hot** par un vecteur à n dimensions dont une seule prend la valeur 1 (toutes les autres valeurs sont nulles).

Cat	one-hot
N	1000
A	0100
V	0010
R	0001

Ce codage qui permet une représentation orthogonale et indépendante des propriétés.

La catégorie A ne partage aucune information avec la catégorie V.

Représentation du sens

Les mots des définitions (et des relations) sont représentés par des *word embeddings* à 50 dimensions.

Ces embeddings ont été calculés en utilisant word2vec à partir du corpus des définitions de GLAWI. Les phrases de ce corpus se composent des définitions précédées du lemme de l'entrée correspondante.

- farine poudre blanche obtenue par la mouture de grains céréaliers et utilisée dans la fabrication du pain et de pâtisseries
- chaise siège avec dossier , sans accoudoirs
- décomposable qui peut être décomposé

taille maximale de la fenêtre = 15 mots ; fréquence minimale des mots = 1 ; sous-échantillonnage = 0.1 ; skip-gram

L'approche est endogène :

- Les définitions sont utilisées pour créer les *word embeddings* des mots des définitions.
- Les définitions sont utilisées comme description du sens des lexèmes cibles et des relations qui caractérisent les cellules cibles.

Expérience 1 : Difficulté de la tâche

Input : {**Cat + Lemme source + Définition**} → Output : {**FAP**}

Input : {**N + LIVRER + celui celle qui livrer de le marchandise**} → Output : {**+eur**}

Exactitude sur le test set

Modèles	Exp. 1
Modèle 'à séquence unique'	47,10 %
Modèle 'à concaténation'	49,35 %

- Identifier les FAP des lexèmes cibles est nettement plus difficile que la résolution de PCFP en flexion.
- Les performances des modèles « à concaténation » sont meilleures (i.e., lorsque les régularités sont identifiées séparément aux 3 niveaux de représentation (formel, catégoriel, sémantique) puis combinées dans un second temps (au niveau du décodeur)).

Exactitude par Cat_{source} et par n. occ.

La difficulté varie en fonction des précédés dérivationnels.

Elle dépend de la concurrence entre les procédés :

- *-able* n'a pas de concurrent (exactitude très forte)
- *-age*, *-ment* et *-ation* sont concurrents (exactitude plus faible)
- la conversion A→V (+er$) est plus rare que la suffixation en *-iser* (exactitude encore plus faible)

Cat_s	Cat_c	FAP	Exp. 1
A	R	+ment$	93 %
A	V	+iser$	**88** %
A	N	+ité$	68 %
A	R	+ement$	84 %
A	V	+er$	**38** %
N	V	+er$	80 %
N	A	+ique$	72 %
N	N	+iste$	67 %
N	N	+eur$	40 %
N	N	+age$	39 %
V	N	+eur$	76 %
V	N	+ement$	**51** %
V	A	+able$	**99** %
V	N	+age$	**48** %
V	N	+ation$	**80** %

Le sens est déterminant... et la forme aussi

Le sens est essentiel dans la prédiction des FAP

Définition	Lem_{src}	Lem_{cib}	$\text{FAP}_{\text{réf}}$	$\text{FAP}_{\text{préd}}$
dégarnir un animal de son corne	corne	écorner	é+er	dé+er
professionnel de le insémination artificiel	insémination	inséminateur	+eur	+iste
qui avoir l aspect de un animal	animal	animaloïde	+oïde	+iforme

Le sens est déterminant... et la forme aussi

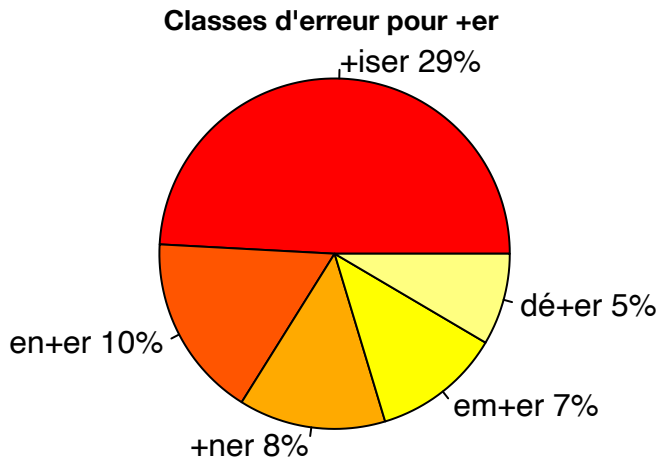
Le radical et la famille dérivationnelle du lemme source contribuent à la sélection du FAP parmi un ensemble de concurrents

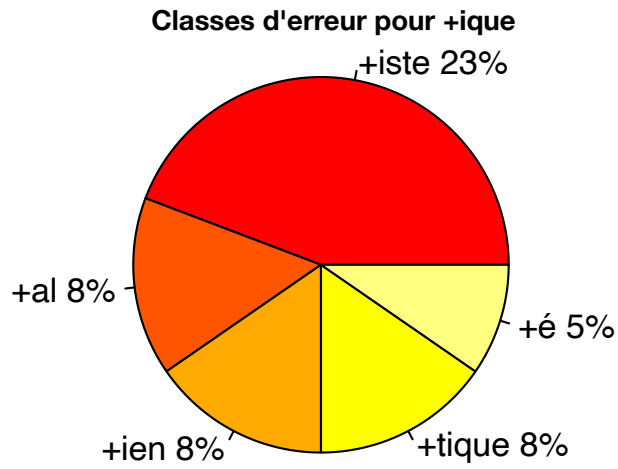
Définition	Lem_{src}	Lem_{cib}	$FAP_{réf}$	$FAP_{préd}$
action de émonder	**émonder**	émondage	+age	+ation
action de taniser ajout de tanin	**taniser**	tanisage	+age	+ation
qui servir de illustration décoratif	**illustration**	illustratif	+atif	+ateur

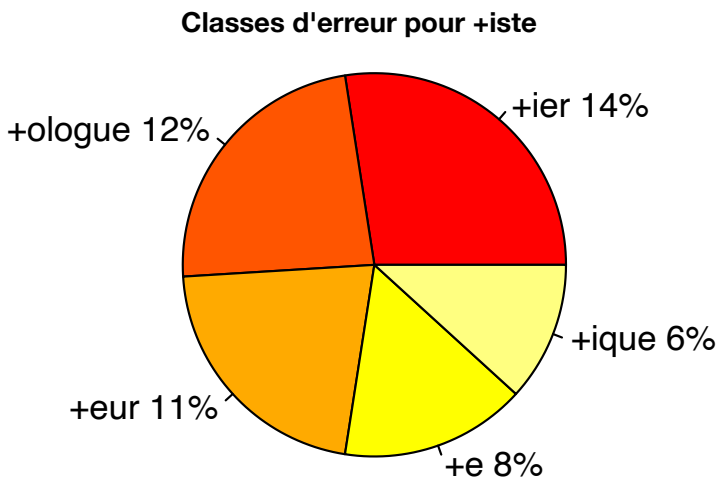
- Plusieurs verbes en *-onder* forment des déverbaux en *-ondation* (*fécondation*, *fondation*, *inondation*, *exondation*) ; plusieurs verbes en é···der ont des noms déverbaux en *-ion* (*élision*, *élucidation*, *érosion*, *évasion*)
- Les noms déverbaux des verbes en *-iser* sont normalement suffixés en *-ation*
- Dans les familles où le nom déverbal à une finale *-ation*, le nom d'instrument est suffixé en *-ateur*

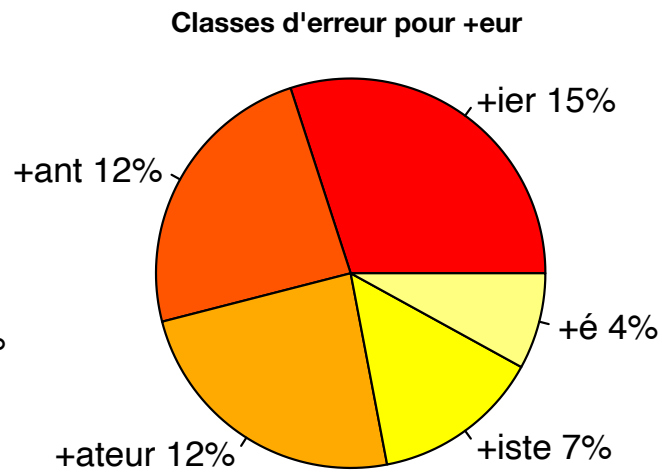
Classes d'erreur

Les erreurs se produisent principalement entre procédés concurrents et entre allomorphes.

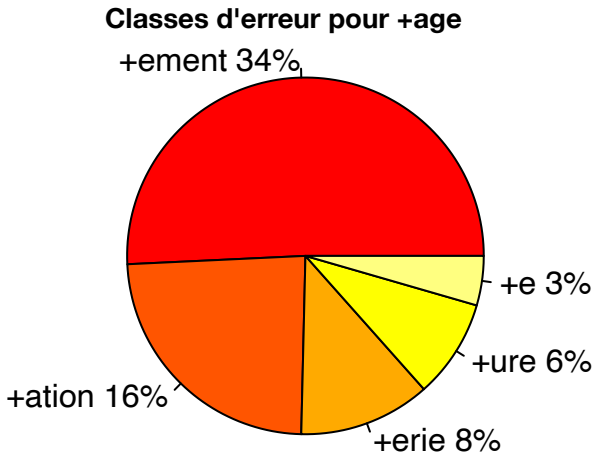


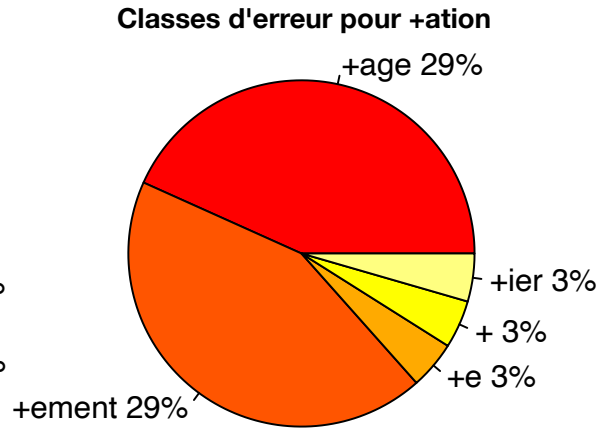


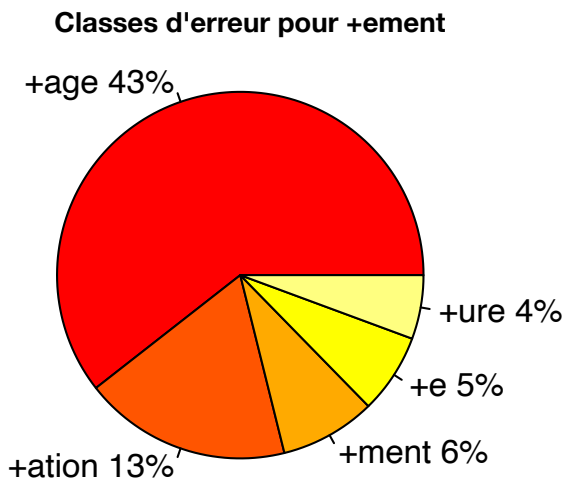




Classes d'erreur







Expérience 2 : Contribution de la forme source

Input : $\left\{ \textbf{Cat} + \textbf{Définition} \right\} \rightarrow$ Output : $\left\{ \textbf{FAP} \right\}$

Input : $\left\{ \textbf{N} + \begin{array}{l} \textbf{celui celle qui livrer} \\ \textbf{de le marchandise} \end{array} \right\} \rightarrow$ Output : $\left\{ \textbf{+eur} \right\}$

Exactitude sur le test set

Modèles	Exp. 1	Exp. 2
M. 'à séquence unique'	47,10 %	31,13 %
M. 'à concaténation'	49,35 %	32,43 %

- Sans le lemme source, il est plus difficile de sélectionner le FAP de référence parmi les procédés concurrents et parmi les allomorphes.
- Le modèle « à concaténation » reste plus performant.

Exactitude par Cat_{source} et par n. occ.

Cat_s	Cat_c	FAP	Exp. 1	Exp. 2
A	R	+ment$	93 %	**70 %**
A	V	+iser$	88 %	64 %
A	N	+ité$	68 %	**26 %**
A	R	+ement$	84 %	31 %
A	V	+er$	38 %	**19 %**
N	V	+er$	80 %	67 %
N	A	+ique$	72 %	66 %
N	N	+iste$	67 %	54 %
N	N	+eur$	40 %	34 %
N	N	+age$	39 %	22 %
V	N	+eur$	76 %	73 %
V	N	+ement$	51 %	29 %
V	A	+able$	99 %	90 %
V	N	+age$	48 %	45 %
V	N	+ation$	80 %	**46 %**

Exactitude par $\text{Cat}_{\text{source}}$ et par n. occ.

Adverbes en *-ment*. Les allomorphes sont déterminés par la forme de l'adjectif au féminin. Le sens seul ne permet pas de prédire ces allomorphes. (L'exactitude pour +ement baisse de 63%)

Définition	Lem_{src}	Lem_{cib}	$\text{FAP}_{\text{réf}}$	$\text{FAP}_{\text{préd}}$
de un manière progressif	progressif	progressivement	+ivement	+ement
de un manière incomplet	incomplet	incomplètement	+ètement	+ement

Noms en *-ation*. Le lemme informe sur la nature savante *vs* vulgaire du radical. (L'exactitude pour +ation baisse de 46%)

Définition	Lem_{src}	Lem_{cib}	$\text{FAP}_{\text{réf}}$	$\text{FAP}_{\text{préd}}$
action de mêler mélange	mêler	mêlement	+ement	+ation

Exactitude par $\text{Cat}_{\text{source}}$ et par n. occ.

Conversion. Le lemme est essentiel pour la prédiction des conversions A→V. (L'exactitude pour +er baisse de 50%)

Définition	Lem_{src}	Lem_{cib}	$\text{FAP}_{\text{réf}}$	$\text{FAP}_{\text{préd}}$
ajouter de le vinaigre balsamique	balsamique	balsamiser ~~vinaigrer~~	+iser	+er

Noms en *-ité*. *-ité* a de nombreux concurrents (*-eur*, *-esse*, *-itude*, *-erie*, *-ise*) et de nombreux allormorphes (*-abilité*). *-ité* construit des noms de qualité. Sa prédiction est pénalisée par le fait que le modèle ne dispose pas de la catégorie A du lexème source. (L'exactitude pour +ité baisse de 42%)

Expérience 3 : Contribution du sens de la source

Input : $\left\{ \textbf{Cat} + \textbf{Lemme source} + \textbf{Relation} \right\} \rightarrow$ Output : $\left\{ \textbf{FAP} \right\}$

Input : $\left\{ \textbf{N} + \textbf{LIVRER} + \begin{array}{l}\textbf{celui celle qui ... de} \\ \textbf{le marchandise}\end{array} \right\} \rightarrow$ Output : $\left\{ \textbf{+eur} \right\}$

Exactitude sur le test set

Modèles	Exp. 1	Exp. 3
M. 'à séquence unique'	47,10 %	47,97 %
M. 'à concaténation'	49,35 %	48,05 %

- Le sens du lexème source **ne contribue pas** à la prédiction du FAP cible.
- Les performances du modèle « à séquence unique » sont améliorées lorsqu'on remplace la description du sens du lexème cible par celle de la relation qui identifie la cellule cible.
- Les performances des deux modèles sont équivalentes.

Exactitude par Cat_{source} et par n. occ.

Les procédés les plus fréquents pour une relation donnée semblent favorisés (*-age*, *-ité*, *-ment*).

Cat_s	Cat_c	FAP	Exp. 1	Exp. 2
A	R	+ment$	93 %	**96** %
A	V	+iser$	88 %	88 %
A	N	+ité$	68 %	**71** %
A	R	+ement$	84 %	81 %
A	V	+er$	38 %	**47** %
N	V	+er$	80 %	**83** %
N	A	+ique$	72 %	**78** %
N	N	+iste$	67 %	67 %
N	N	+eur$	40 %	**53** %
N	N	+age$	39 %	**34** %
V	N	+eur$	76 %	76 %
V	N	+ement$	51 %	42 %
V	A	+able$	99 %	96 %
V	N	+age$	48 %	**63** %
V	N	+ation$	80 %	74 %

Ouvrir la boîte

L'omission du sens du lexème source dans la définition modifie les connaissances du système.

L'information sémantique devient une relation du type :

(action | résultat | maniére | technique) de

Nous avons comparé la représentation interne du système (la matrice des états d'activation, la grande flèche orange à l'extrémité de l'encodeur) des modèle « à séquence unique » des expériences 1 et 3 pour observer ces changements.

Ouvrir la boîte

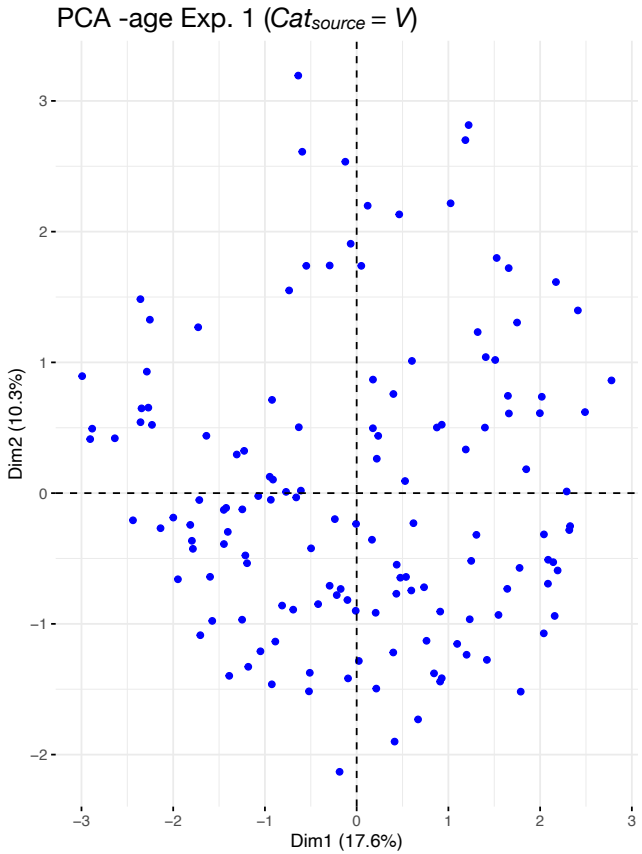
La comparaison porte sur deux sous-ensembles du jeu d'entraînement

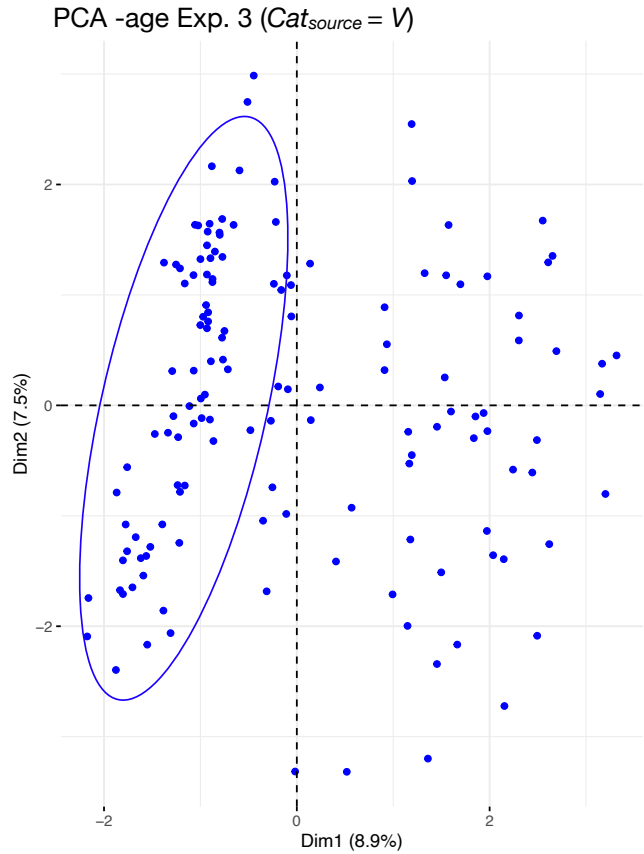
1. les 142 entrées dont le lexème source est un verbe, dont le lexème cible est un nom et dont le FAP du lexème cible est +age$ (ex. **laver-lavage**).
2. les 406 entrées dont le lexème source est un verbe, dont le lexème cible est un nom et dont le FAP du lexème cible est +age$, +ement$ ou +ation$.

Méthode

- Compilation des matrices des dérivations à observer dans les modèles des expériences 1 et 3.
  - Matrice 142 $\times$ 300 pour la dérivation en **-age**.
  - Matrice 406 $\times$ 300 pour les dérivations en **-age**, **-ement** et **-ation**.
- Pour chaque sous-ensemble, application d'une PCA à chacune des deux matrices pour observer la distribution des représentations.

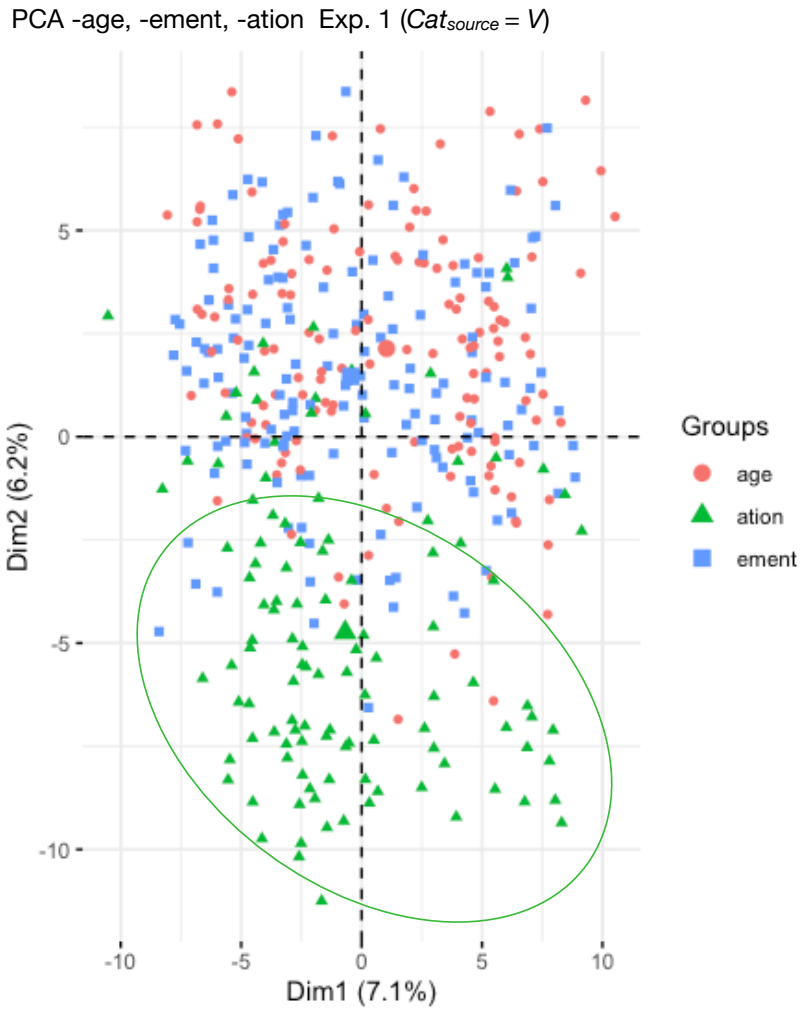
Ouvrir la boîte

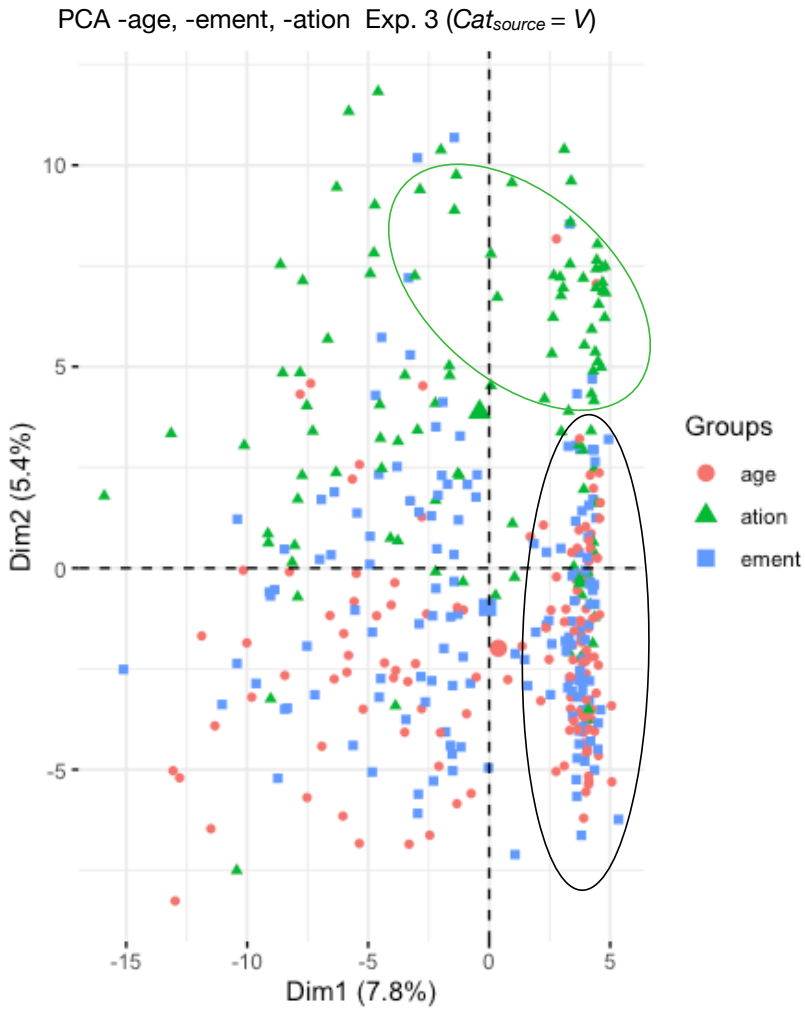




- Les représentations construites à partir des relations sémantiques (exp. 3) sont sont spatialement plus proches et donc structurellement plus similaires que celles construites à partir de la définition du lexème cible (exp. 1).

Ouvrir la boîte





Ouvrir la boîte

- Le modèle de l'expérience 1 montre que les dérivations en **-ation** se distinguent de celles en **-age** et **-ement** (probablement parce que *-ation* préfère les radicaux savants).
- L'expérience 3 fait apparaître la même distinction
- La similarité des représentations dans le modèle de l'expérience 3 concerne les mêmes dimensions pour les trois suffixations.

Robustesse sémantique

Afin d'estimer la sensibilité du système à la variabilité des descriptions sémantiques, nous avons testé les modèles en remplaçant les définitions GLAWI par celles du TLFi (Pierrel et al., 2004).

Nous avons créé un nouveau jeu de données en sélectionnant les entrées du notre jeu de test dont les définitions dans le TLFi sont des définitions morphologiques différentes de celles de GLAWI et qui ont une taille est comprise entre 2 et 15 mots.

Le jeu de test TLFi comporte 2326 entrées

Définitions du nom **possibilité** :

GLAWI chose possible

TLFi faire de être possible qualité de ce qui être possible

Robustesse sémantique

Exactitude des modèles de l'expérience 1 sur le jeu de test TLFi

Modèles	test TLFi
Modèle 'à séquence unique'	49,50 %
Modèle 'à concaténation'	50,60 %

- Les modèles disposent de représentations des descriptions sémantiques qui captent correctement le sens du lexème cible.
- L'exactitude est améliorée lorsque l'on utilise les définitions du TLFI (effet de la longueur des définitions ?)
- Le modèle « à concaténation » obtient une exactitude plus élevée que le modèle « à séquence unique ».

Prédire les lexèmes de la famille dérivationnelle

Caractérisation formelle des relations dérivationnelles

Le lexique Glawinette

Questions

Méthodes et données

Expériences

Conclusion

Conclusion

Nous avons construit deux modèles (« à concaténation » et « à séquence unique ») pour traiter une version simplifiée de PCFP en morphologie dérivationnelle.

Les entrée des modèles se composent de trois types d'informations : **$\text{Cat}_{\text{cible}}$**, **$\text{Lemme}_{\text{source}}$**, **$\text{Définition}_{\text{cible}}$**. La sortie est le **$\text{FAP}_{\text{cible}}$**.
L'exactitude atteint 49,35% pour le modèle « à concaténation » de l'expérience 1.

Les erreurs de prédiction sont généralement morphologiquement "valides".

Le **lemme source** détermine fortement la prédiction du FAP (expérience 2).

L'omission du sens du lexème source ne dégrade pas la prédiction du FAP (expérience 3).

Perspectives

Poursuivre l'analyse des résultats et des erreurs

Ajouter la catégorie du lexème source aux entrées des modèles

Considérer les définitions de plus de 15 mots

Refaire les mêmes expériences en anglais (lexique **Englawinette**) et en italien (lexique **GlawITina)**

À plus long terme, nous envisageons de comparer les sorties du système aux réponses de locuteurs auxquels on présenterait les mêmes stimuli.

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*, 54–81. Oxford: Oxford University Press.

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.

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

Boyé, Gilles & Gauvain Schalchli. 2019. Realistic data and paradigms: The Paradigms Cell Finding Problem. *Morphology* 29(2). 199–248.

Calderone, Basilio, Nabil Hathout & Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. In *Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, 196–204.

Cotterell, Ryan, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov & David Yarowsky. 2017. Paradigm completion for derivational morphology. In *Proceedings of the 2017 conference on empirical methods in natural language processing, EMNLP 2017*, 714–720. Copenhagen, Denmark.

Fradin, Bernard. 2020. Characterizing derivational paradigms. In Jesús Fernández-Domínguez, Alexandra Bagasheva & Cristina Lara-Clares (eds.), *Paradigmatic relations in word formation*, 49–84. Brill.

Fradin, Bernard. 2021. Caractériser les paradigmes dérivationnels. *Verbum* XLIII(1). 149–178.

Goldsmith, John. 2001. Unsupervised learning of the morphology of natural language. *Computational Linguistics* 27(2). 153–198.

Références

Goldsmith, John. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12(4). 353–371.

Hathout, Nabil. 2009. *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. 2011. Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*. In *Des unités morphologiques au lexique*, 251–318. Paris: Hermès Science-Lavoisier.

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.

Références

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`.

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

Hathout, Nabil & Fiammetta Namer. 2022. ParaDis: a family and paradigm model. *Morphology* 32(2). 153–195.

Références

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.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8). 1735–1780.

Malouf, Rob. 2016. Generating morphological paradigms with a recurrent neural network. *San Diego Linguistic Papers* 6. 122–129.

Malouf, Rob. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology* 27(4). 431–458.

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

Références

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 26e conférence annuelle sur le traitement automatique des langues naturelles (taln 2019)*, 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.

Pierrel, Jean-Marie, Jacques Dendien & Pascale Bernard. 2004. Le TLFi ou Trésor de la Langue Française informatisé. In Geoffrey Williams & Sandra Vessier (eds.), *Proceedings of the 11th EURALEX international congress*, 165–170. Lorient, France.

Roché, Michel. 2017. Les familles dérivationnelles : comment ça marche ? Manuscrit.

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 & 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.

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