Demonette : une base de données morphologique du français

Nabil Hathout

CLLE/ERSS, CNRS

Séminaire CARTEL / Master LITL
22 octobre 2018

Vue d'ensemble

Demonette et Demonext...

Demonette

- Demonette = **DE**rivational **MO**rphological **NET**work
- Demonette est une base de données morphologique du français
- Collaboration entre **Fiammetta Namer** (Université de Lorraine) et **Nabil Hathout** (CNRS)
- Plusieurs versions de Demonette ont déjà été créées et diffusées (v1.0 à v1.3)

Demonette et Demonext... (2)

Demonext

- Demonext = **NEXT DEMON**ette
- Demonext vise à créer la prochaine version de Demonette (v2)
- Projet financé par l'ANR (Agence Nationale de la Recherche)
- Le consortium est formé de l'ATILF (Nancy), CLLE (Toulouse), STL (Lille) et LLF (Paris)
  - Il réunit des spécialistes de description morphologique, de traitement automatique des langues, de didactique du français et des orthophonistes.

- **Abus de langage :** on identifie le projet et la base de donnée. **Demonette** sert à désigner l'ensemble.

Demonette

Objectifs

- Demonette vise à construire une base de donnée morphologique
  - à large couverture
  - comportant des descriptions riches et fiables
- Demonette répondra à des besoins multiples
  - confirmation empirique et élaboration d'hypothèses en morphologie
  - développement d'outils de TAL
  - enseignement du vocabulaire (primaire) et de la morphologie (université)
  - traitement des troubles du langage développementaux ou acquis

Ressource collaborative

- Les morphologues décrivent régulièrement leurs analyses dans des bases de données (thèses, articles de recherche).
  - Les descriptions sont riches et fiables (évaluation par le jury de thèse et par les relectures des revues).
- Demonette est un cadre où toutes ces bases de données peuvent être intégrées, et rendues disponibles (licence CC) et exploitables.
- À termes, Demonette permettra au morphologue de :
  - récupérer rapidement un ensemble conséquent d'analyses d'un procédé ou d'un phénomène qui l'intéresse
  - rendre public et diffuser ses analyses
  - pourra être utilisé comme un service =
    décrire et réaliser les analyses directement dans l'outil

Ressource intégrées à Demonette

Demonette v1

- v1.0 à v1.2 :
  analyses **DeriF** des entrées du TLFi (Fiammetta Namer)
  base **Morphonette** (Nabil Hathout)
- v1.3 : + base **Lexeur** (Cécile Fabre et Nabil Hathout)

Ressource intégrées à Demonette (2)

Demonette v2

- Base des **Convers** de Delphine Tribout (thèse : « Les conversions de nom à verbe et de verbe à nom en français », 2010)
- Base **DeNom** de Jana Strnadova (thèse : « Les réseaux adjectivaux. Sur la grammaire des adjectifs dénominaux en français », 2014)
- Base **MORDAN** de Aurore Koehl (thèse : « La construction morphologique des noms désadjectivaux suffixés en français », 2012)
- Base **DiMoC** de Michel Roché (articles de recherche sur *-ier*, *-at*, *-eraie* mais aussi *-isme*, *-iste*)
- Base **Lexeur** de Cécile Fabre et Nabil Hathout (2002)
- + 2 ressources non (totalement) révisées manuellement : analyses **DeriF** des entrées du TLFi et **Morphonette**

Petite démo

https://demonette.atilf.fr/

démonstration, *dissertation*, *adoucir*

Vue d'ensemble

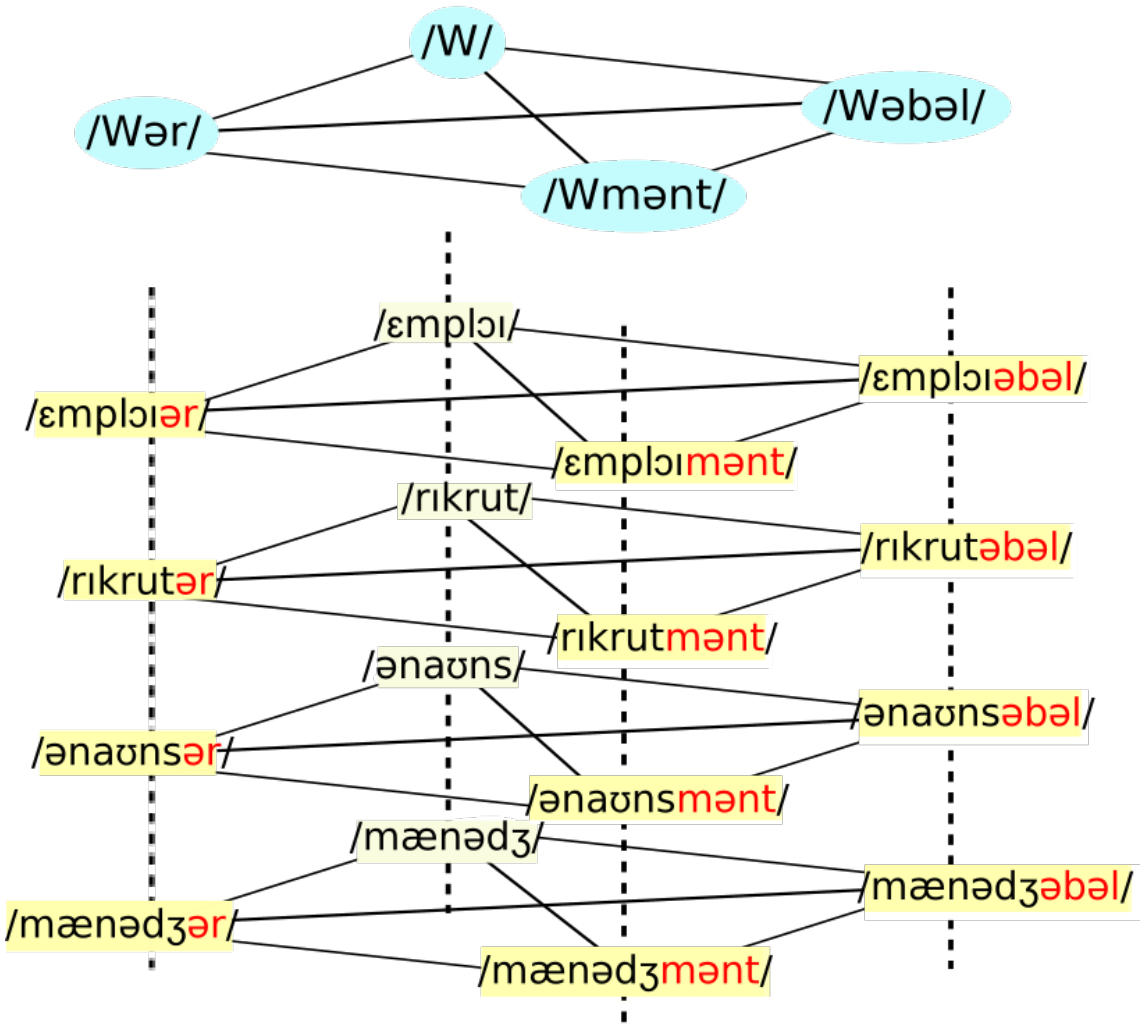
Demonette et ParaDis

- Demonette et ParaDis sont les faces appliquée et théorique du même projet scientifique (Fiammetta Namer et Nabil Hathout)
- ParaDis = **PARADIgms versus DIScrepancies**
  « les paradigmes contre les décalages »
- La description paradigmatique de la morphologie dérivationnelle permet d'analyser naturellement un grand nombre de phénomènes non canoniques comme
  - les **décalages** (*lieu* → *localiser*)
  - la **surabondance** (*religion* → *interreligion*, *interreligieux*)
  - la **supplétion** (*eau* → *aqueux*, *hydrolique*)
  - les **syncrétismes** (*Italie* → *italien* (gentilé), *Italie* → *italien* (langue))
- ParaDis est un modèle théorique dans lequel les relations morphologiques s'organisent en **paradigmes**.

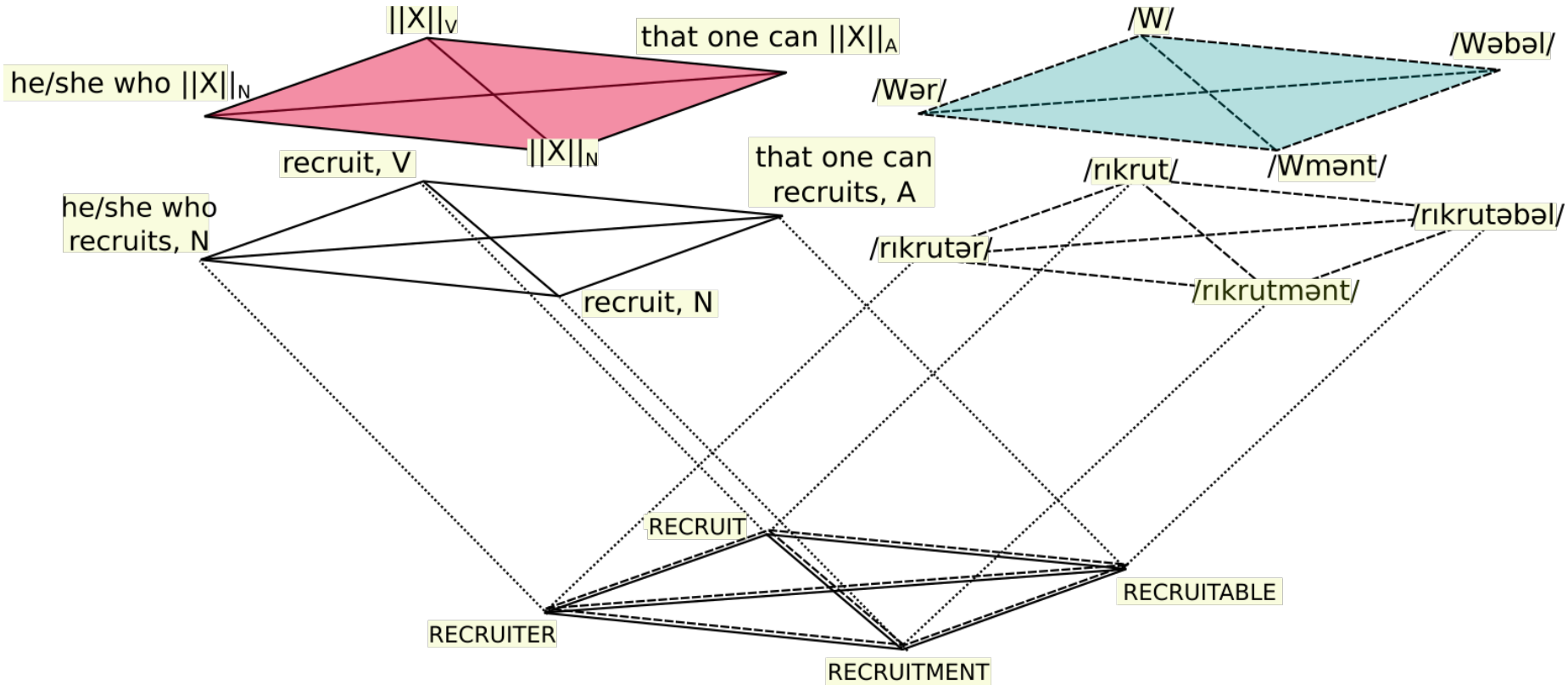
ParaDis

- ParaDis s'inscrit dans un cadre Familles et Paradigmes
- Dans une perspective lexématique,
  - la dérivation est une **relation directe** entre un lexème base et un lexème dérivé (*démontrer* $\rightarrow$ *démonstration*)
  - les **relations indirectes** (*prédateur* $\rightarrow$ *prédation*) ne sont pas analysables. Pas de base $^{\circ}$*préder* en français.
- L'analyse morphologique consiste à décrire la structure des familles morphologiques et la façon dont elle se superposent pour former des paradigmes
- la notion de famille et celle de paradigme sont généralisées aux niveaux de description formel, sémantique et morphologique

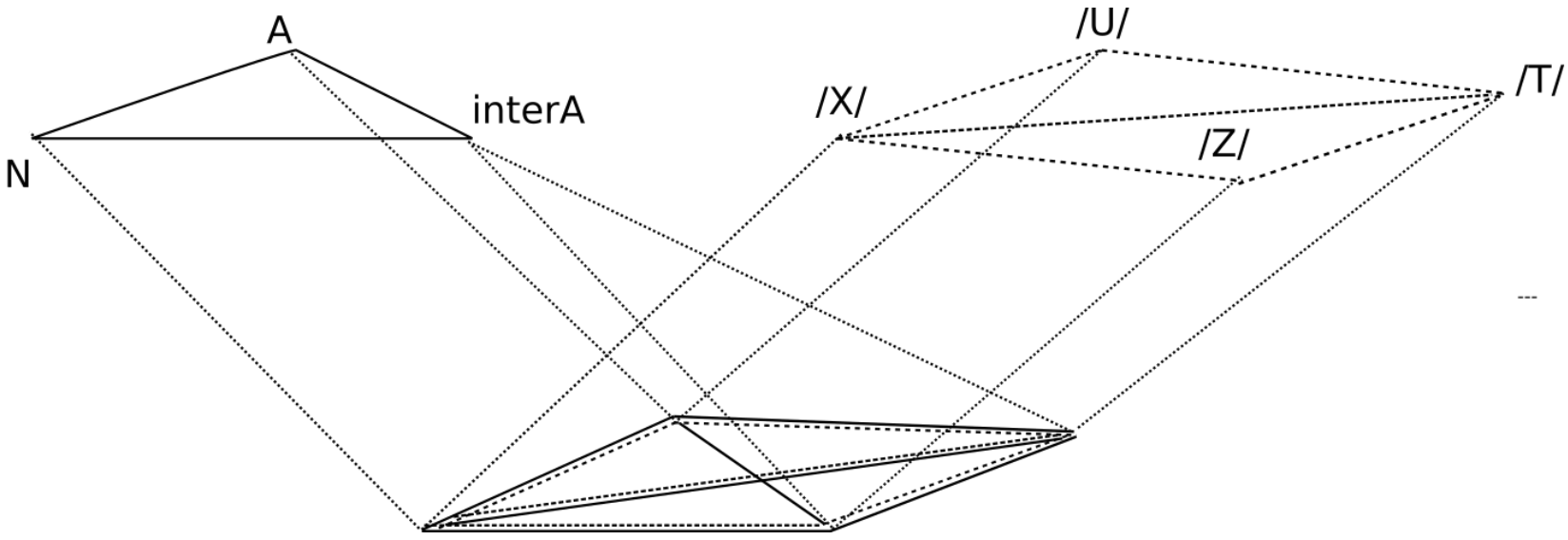
Paradigme formel



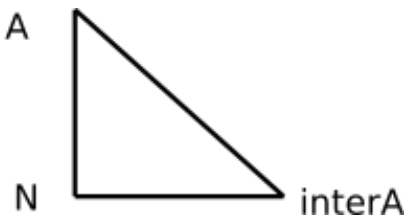
Paradigme dérivationnel



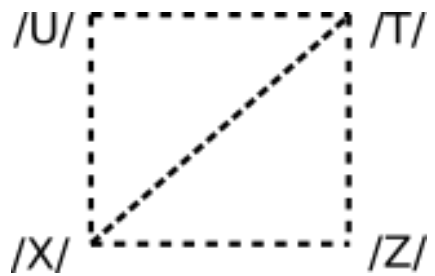
Paradigme dérivationnel (2)



N:	$\|\|Y\|\|$, N
A:	$\|\|Y\|\|$, A
interA:	between several $\|\|Y\|\|$, A



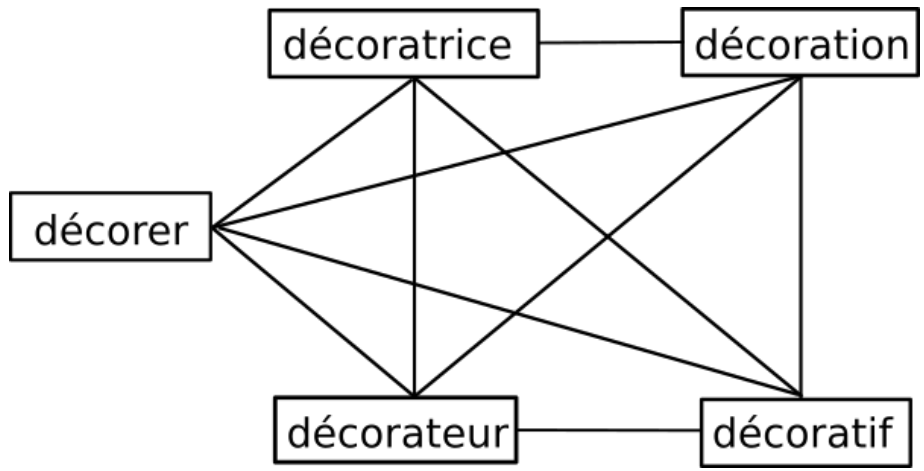
X:	$\exists W,\ /X/=/W/$
U:	$\exists W,\ /U/=/W\text{ik}/\vee/W\text{əl}/\vee/W\text{əs}/\vee/W\text{ər}/$
Z:	$\exists W,\ /Z/=/\text{ɪntər}W/$
T:	$\exists V,\ /T/=/\text{ɪntər}V/\wedge$
	$\exists W,(/V/=/W\text{ik}/\vee/W\text{əl}/\vee/W\text{əs}/\vee/W\text{ər}/)$



Vue d'ensemble

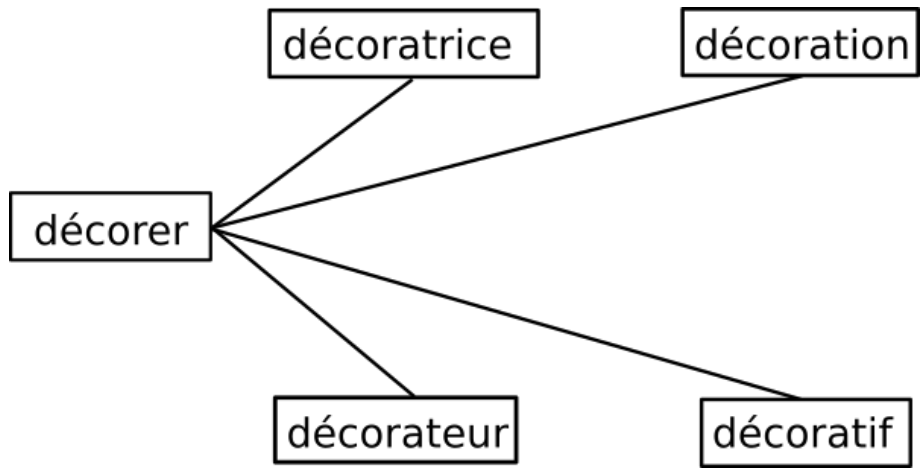
Demonette

- Demonette décrit des familles morphologiques
- Une **famille morphologique** est un graphe connexe
- Dans une famille, les sommets sont des lexèmes et les arcs sont des relations dérivationnelles
- Demonette contient des lexèmes simples et des lexèmes construits



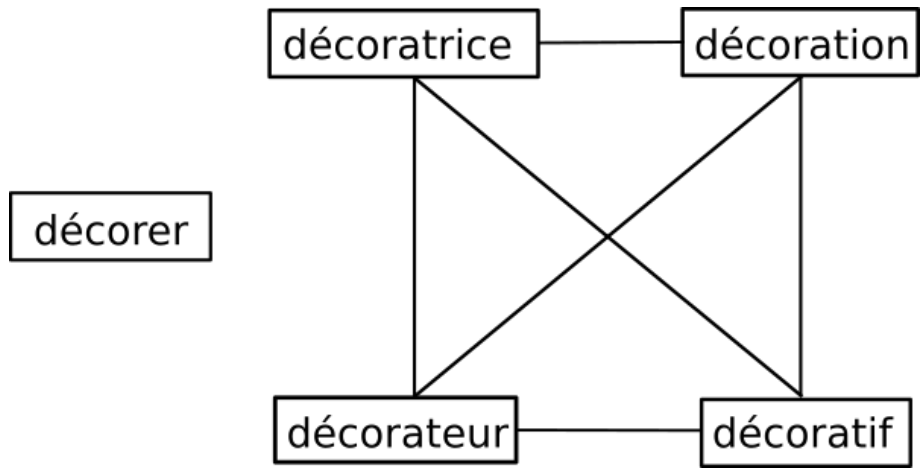
Relations directes et indirectes

- Les relations **directes** connectent les dérivés à leurs bases
- Les relations **indirectes** s'établissent entre les mots de la famille dérivationnelle qui sont interprédictibles sémantiquement



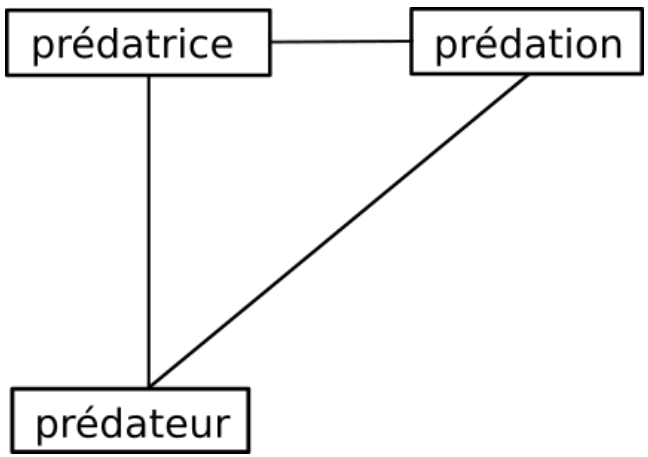
Relations directes et indirectes

- Les relations **directes** connectent les dérivés à leurs bases
- Les relations **indirectes** s'établissent entre les mots de la famille dérivationnelle qui sont interprédictibles sémantiquement



Relations directes et indirectes

- Les relations **directes** connectent les dérivés à leurs bases
- Les relations **indirectes** s'établissent entre les mots de la famille dérivationnelle qui sont interprédictibles sémantiquement

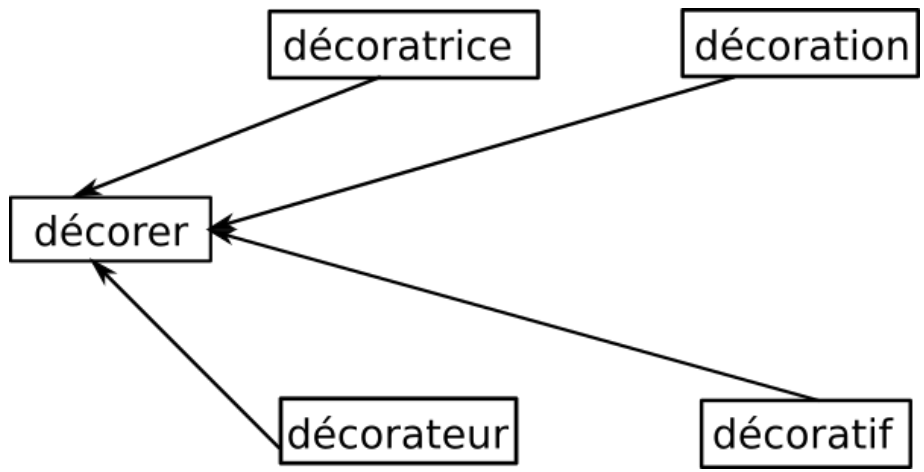


Relations directes et indirectes

- Les familles sont des graphes orientés
  - $W_2 \leftarrow W_1$ décrit la **motivation morphologique** de W_2 relativement à W_1
  - La motivation est symétrique pour la plupart des couples de lexèmes
- Les relations directes **descendantes** connectent un dérivé à sa base ou à un ascendant plus distant
- Les relations directes **ascendantes** connectent un mot à l'un de ses dérivés ou des ses descendants distants
- Les relations **indirectes** sont bi-orientées

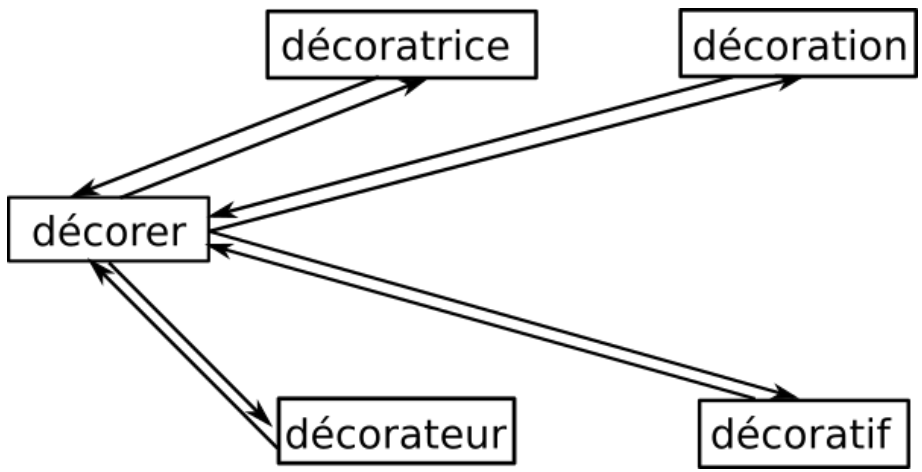
Relations directes et indirectes

- Les familles sont des graphes orientés
  - $W_2 \leftarrow W_1$ décrit la **motivation morphologique** de W_2 relativement à W_1
  - La motivation est symétrique pour la plupart des couples de lexèmes
- Les relations directes **descendantes** connectent un dérivé à sa base ou à un ascendant plus distant
- Les relations directes **ascendantes** connectent un mot à l'un de ses dérivés ou des ses descendants distants
- Les relations **indirectes** sont bi-orientées



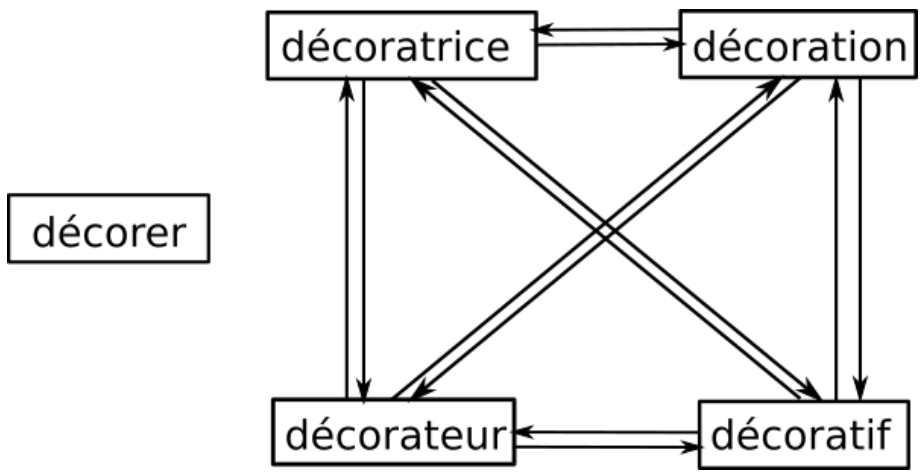
Relations directes et indirectes

- Les familles sont des graphes orientés
  - $W_2 \leftarrow W_1$ décrit la **motivation morphologique** de W_2 relativement à W_1
  - La motivation est symétrique pour la plupart des couples de lexèmes
- Les relations directes **descendantes** connectent un dérivé à sa base ou à un ascendant plus distant
- Les relations directes **ascendantes** connectent un mot à l'un de ses dérivés ou des ses descendants distants
- Les relations **indirectes** sont bi-orientées



Relations directes et indirectes

- Les familles sont des graphes orientés
  - $W_2 \leftarrow W_1$ décrit la **motivation morphologique** de W_2 relativement à W_1
  - La motivation est symétrique pour la plupart des couples de lexèmes
- Les relations directes **descendantes** connectent un dérivé à sa base ou à un ascendant plus distant
- Les relations directes **ascendantes** connectent un mot à l'un de ses dérivés ou des ses descendants distants
- Les relations **indirectes** sont bi-orientées



Complexité

Demonette contient des relations de différentes complexité

- Une relation **directe** est **simple** lorsqu'elle connecte un dérivé et sa base (= une étape de dérivation)
- Une relation **indirecte** $W_2 \leftarrow W_1$ est considérée comme **simple**
  - si W_1 et W_2 sont connectées par un chemin de deux relations simples, ou
  - si W_1 et W_2 appartiennent à une série de relations simples
- Une relation est **lexicale** si elle s'établit entre des lexèmes morphologiquement apparentés, mais qui ne sont pas dans une relation dérivationnelle régulière.
- Toutes les autres relations sont **complexes**

Complexité

Demonette contient des relations de différentes complexité

- Une relation **directe** est **simple** lorsqu'elle connecte un dérivé et sa base (= une étape de dérivation)
- Une relation **indirecte** $W_2 \leftarrow W_1$ est considérée comme **simple**
  - si W_1 et W_2 sont connectées par un chemin de deux relations simples, ou
  - si W_1 et W_2 appartiennent à une série de relations simples
- Une relation est **lexicale** si elle s'établit entre des lexèmes morphologiquement apparentés, mais qui ne sont pas dans une relation dérivationnelle régulière.
- Toutes les autres relations sont **complexes**

entrée	**complexité**	**orientiation**	
chanteur $\leftarrow$ chanter	simple	direct	descendant
chanter $\leftarrow$ chanteur	simple	direct	ascendant

Complexité

Demonette contient des relations de différentes complexité

- Une relation **directe** est **simple** lorsqu'elle connecte un dérivé et sa base (= une étape de dérivation)
- Une relation **indirecte** $W_2 \leftarrow W_1$ est considérée comme **simple**
  - si W_1 et W_2 sont connectées par un chemin de deux relations simples, ou
  - si W_1 et W_2 appartiennent à une série de relations simples
- Une relation est **lexicale** si elle s'établit entre des lexèmes morphologiquement apparentés, mais qui ne sont pas dans une relation dérivationnelle régulière.
- Toutes les autres relations sont **complexes**

entrée	complexité	orientiation	
chanteur ← chanteuse	simple	indirect	–
prédateur ← prédation	simple	indirect	–

Complexité

Demonette contient des relations de différentes complexité

- Une relation **directe** est **simple** lorsqu'elle connecte un dérivé et sa base (= une étape de dérivation)
- Une relation **indirecte** $W_2 \leftarrow W_1$ est considérée comme **simple**
  - si W_1 et W_2 sont connectées par un chemin de deux relations simples, ou
  - si W_1 et W_2 appartiennent à une série de relations simples
- Une relation est **lexicale** si elle s'établit entre des lexèmes morphologiquement apparentés, mais qui ne sont pas dans une relation dérivationnelle régulière.
- Toutes les autres relations sont **complexes**

entrée	complexité	orientiation	
interrogatoire ← interroger	lexical	–	–
mensonge ← mentir	lexical	–	–

Complexité

Demonette contient des relations de différentes complexité

- Une relation **directe** est **simple** lorsqu'elle connecte un dérivé et sa base (= une étape de dérivation)
- Une relation **indirecte** $W_2 \leftarrow W_1$ est considérée comme **simple**
  - si W_1 et W_2 sont connectées par un chemin de deux relations simples, ou
  - si W_1 et W_2 appartiennent à une série de relations simples
- Une relation est **lexicale** si elle s'établit entre des lexèmes morphologiquement apparentés, mais qui ne sont pas dans une relation dérivationnelle régulière.
- Toutes les autres relations sont **complexes**

entrée	**complexité**	**orientiation**	
progressivité ← progresser	complex	direct	descendant
progressivité ← progression	complex	indirect	–

Description fine

- Demonette décrit l'ensemble des arcs qui composent les familles dérivationnelles
- Une relation dérivationnelle est identifiée au couple de lexèmes qu'elle connecte
- Demonette contient la **forme de citation** des lexèmes et leurs **catégories grammaticales**
- En plus des propriétés d'orientation et de complexité, les relations sont caractérisées par :
  - le **type** de la construction (*pref*, *suf*, *conv*)
  - l'**exposant**

Description fine

- Demonette décrit l'ensemble des arcs qui composent les familles dérivationnelles
- Une relation dérivationnelle est identifiée au couple de lexèmes qu'elle connecte
- Demonette contient la **forme de citation** des lexèmes et leurs **catégories grammaticales**
- En plus des propriétés d'orientation et de complexité, les relations sont caractérisées par :
  - le **type** de la construction (*pref*, *suf*, *conv*)
  - l'**exposant**

entrée	**form$_1$**	**cat$_1$**	**form$_2$**	**cat$_2$**
production ← produire	production	Ncms	produire	Vmn----
progessivité ← progessif	progressivité	Ncfs	progressif	Afpms

Description fine

- Demonette décrit l'ensemble des arcs qui composent les familles dérivationnelles
- Une relation dérivationnelle est identifiée au couple de lexèmes qu'elle connecte
- Demonette contient la **forme de citation** des lexèmes et leurs **catégories grammaticales**
- En plus des propriétés d'orientation et de complexité, les relations sont caractérisées par :
  - le **type** de la construction (*pref*, *suf*, *conv*)
  - l'**exposant**

entrée	**typ$_1$**	**exp$_1$**	**typ$_2$**	**exp$_2$**
conceptrice ← concevoir	suf	rice	–	–
hydroplaner ← hydroplanage	–	–	suf	age
analyser ← analyse	conv	–	conv	–

Description fine (2)

- Chaque lexème est muni d'un **type morpho-sémantique**
  @, @ACT, @RES, @AGM, @AGF, @PRP
- La motivation sémantique de W_2 relativement à W_1 est décrite au moyen de deux **gloses**, l'une **concrète** et l'autre **abstraite**
- Le sens des lexèmes est **cumulatif**
  - Le sens morphologique d'un mot est une agrégation des sens élémentaires redondants
  - Chaque relation dérivationnelle contribue au sens des lexèmes qu'elle connecte

@	@ACT	@RES	@AGM	@AGF	@PRP
décorer	décoration	décoration	décorateur	décoratrice	décoratif
aboyer	aboiement	aboiement	aboyeur	aboyeuse	
	audition	audition	auditeur	auditrice	auditif

Description fine (2)

- Chaque lexème est muni d'un **type morpho-sémantique**
  @, @ACT, @RES, @AGM, @AGF, @PRP
- La motivation sémantique de W_2 relativement à W_1 est décrite au moyen de deux **gloses**, l'une **concrète** et l'autre **abstraite**
- Le sens des lexèmes est **cumulatif**
  - Le sens morphologique d'un mot est une agrégation des sens élémentaires redondants
  - Chaque relation dérivationnelle contribue au sens des lexèmes qu'elle connecte

entrée	**glose concrète**
transporteur ← transporter	(masc agent OR instrument) of transporter

Description fine (2)

- Chaque lexème est muni d'un **type morpho-sémantique**
  @, @ACT, @RES, @AGM, @AGF, @PRP
- La motivation sémantique de W_2 relativement à W_1 est décrite au moyen de deux **gloses**, l'une **concrète** et l'autre **abstraite**
- Le sens des lexèmes est **cumulatif**
  - Le sens morphologique d'un mot est une agrégation des sens élémentaires redondants
  - Chaque relation dérivationnelle contribue au sens des lexèmes qu'elle connecte

entrée	**glose abstraite**
transporteur ← transporter	(masc agent OR instrument) of @

Description fine (2)

- Chaque lexème est muni d'un **type morpho-sémantique**
  @, @ACT, @RES, @AGM, @AGF, @PRP
- La motivation sémantique de W_2 relativement à W_1 est décrite au moyen de deux **gloses**, l'une **concrète** et l'autre **abstraite**
- Le sens des lexèmes est **cumulatif**
  - Le sens morphologique d'un mot est une agrégation des sens élémentaires redondants
  - Chaque relation dérivationnelle contribue au sens des lexèmes qu'elle connecte

entrée	**glose concrète**
transporteur ⟵ transporter	(masc agent OR instrument) of transporter
transporteur ⟵ transporteuse	he whose fem correspondant is the transporteuse
transporteur ⟵ transport	(masc agent OR instrument) of the transport

Description fine (2)

- Chaque lexème est muni d'un **type morpho-sémantique**
  @, @ACT, @RES, @AGM, @AGF, @PRP
- La motivation sémantique de W_2 relativement à W_1 est décrite au moyen de deux **gloses**, l'une **concrète** et l'autre **abstraite**
- Le sens des lexèmes est **cumulatif**
  - Le sens morphologique d'un mot est une agrégation des sens élémentaires redondants
  - Chaque relation dérivationnelle contribue au sens des lexèmes qu'elle connecte

entrée	**glose abstraite**
transporteur ← transporter	(masc agent OR instrument) of @
transporteur ← transporteuse	he whose fem correspondant is the @AGF
transporteur ← transport	(masc agent OR instrument) of the @ACT

Taille, origine et disponibilité

- Demonette v1.3: 167369 couples de lexèmes ;
  DériF + Morphonette + Verbaction + Lexeur.
- La **source** de chaque élément d'information est renseignée
- Demonette est distribuée sous licence Creative Commons.
- Le format peut être étendu pour décrire des types d'information qui n'ont pas encore été rencontrés

Form_1	doseuse	Process_1	suf
Source_Form_1	tlfnome	Exponent_1	euse
Form_2	dosage	Source_Constr_1	demonette
Source_Form_1	tlfnome	Process_2	suf
Cat_1	Ncfs	Exponent_2	age
Source_Cat_1	tlfnome	Source_Constr_2	demonette
Cat_2	Ncms	Type_1	@AGF
Source_Cat_2	tlfnome	Source_Type_1	demonette
Complexity	simple	Type_1	@RES
Source_Complexity	demonette	Source_Type_1	demonette

Vue d'ensemble

Alimentation concurrente de la base

- Demonette est construit par plusieurs équipes qui travaillent sur des parties différentes de la base
- **Solution envisagée initalement :**
  - utiliser un wiki pour permettre la modification en temps réel de la base
- **Problème de cohérence :**
  - Comment gérer les interférences entre les traitements des différentes parties par les éditeurs ?
  - Certaines opérations seront réalisées par lots.

Alimentation concurrente de la base

- Demonette est construit par plusieurs équipes qui travaillent sur des parties différentes de la base
- **Solution envisagée initalement :**
  - utiliser un wiki pour permettre la modification en temps réel de la base
- **Problème de cohérence :**
  - Comment gérer les interférences entre les traitements des différentes parties par les éditeurs ?
  - Certaines opérations seront réalisées par lots.
- **Solution retenue :**
  - utiliser git pour résoudre les conflits entre les contributions des différentes équipes.
  - git est un système de gestion de versions décentralisé créé par Linus Torvalds pour le suivi des modifications et l'archivage des versions du noyau linux.
  - 3 types d'utilisateurs : éditeurs, responsables de tâche, administrateurs

Nouvelle description sémantique

- Demonette v1.3 ne contient que 6 types morpho-sémantiques et une cinquantaine de gloses abstraites différentes.
- Les familles de Demonette v1.3 sont constituées de relations simples, bien définies, organisées autour d'un prédicat verbal (@)
- Ce mode de description ne permet pas un passage à l'échelle.
- Thèse de **Daniele Sanacore**

Éléments de solution 1

remplacer les définitions par des énoncés définitoires où les deux lexèmes sont présents :

	W_2	glose concrète
v1.3	ramasseur	'celui qui ramasse'
v2	ramasseur	'un ramasseur ramasse'

Nouvelle description sémantique (2)

Éléments de solution 2

Décomposer les types morpho-sémantiques pour distinguer les propriétés **ontologiques** des propriétés **argumentales**.

W_1	W_2	type W_1	type W_2
ramasser	ramasseur	@	@AGM

W_1	W_2	type W_1	type W_2	relation	rôle W_2
ramasser	ramasseur	DYN	HUM	PRED:ARG	AGENT

Nouvelle description sémantique (3)

Éléments de solution 3

Ajouter une description sémantique pour l'ensemble de la famille dérivationnelle au moyen des **frames sémantiques** de **FrameNet**.

- connecter chaque membre de la famille à un élément de la définition et des autres sections du frame (core et non-core).
- dans quelle mesure peut-on utiliser les descriptions existantes ?
- peut-on utiliser le FrameNet de l'anglais ?
- faut-il décrire une famille morphologique au moyen d'un seul frame ou la recouvrir par plusieurs frames ?
- doit-on concevoir un jeu de frames spécifiques pour les familles morphologiques du français ?

Description phonologiques actuelle

- transcription phonétique des radicaux et des affixes
- syllabification
- taille du radical
- propriétés de la dernière syllabe à la jointure du radical et de l'exposant
- variation formelle entre W_1 et W_2

W_1	**Phon**$_1$	**Affix**$_1$	W_2	**Phon**$_2$	**Affix**$_2$
loueur	lu.œr	œr	location	lɔ.ka.sjɔ̃	jɔ̃
louer	lu		loueur	lu.œr	œr

Description phonologiques actuelle

- transcription phonétique des radicaux et des affixes
- syllabification
- taille du radical
- propriétés de la dernière syllabe à la jointure du radical et de l'exposant
- variation formelle entre W_1 et W_2

W_1	**Stem**$_i$	**Size**$_i$	**LastOns**$_i$	**LastNucl**$_i$	**LastCod**$_i$
loueur	lu	1	l	u	
location	lɔ.kas	2	k	a	s

Description phonologiques actuelle

- transcription phonétique des radicaux et des affixes
- syllabification
- taille du radical
- propriétés de la dernière syllabe à la jointure du radical et de l'exposant
- variation formelle entre W_1 et W_2

W_1	**Stem$_1$**	W_2	**Stem$_2$**	**Stem Variation**
loueur	lu	location	lɔkas	lu/lɔkas
location	lɔkas	locatif	lɔkat	s/t
louer	lu	loueur	lu	=

Comment décrire les propriétés phonologiques dans Demonette v2 ?

Quelles formes décrire ?

Les relations entre les formes d'un lexème peuvent être décrites :

1. en énumérant toutes les formes fléchies de W_1 et de W_2 ; cela impose de décrire les relations phonétiques entre tous les couples de formes de W_1 et de W_2
2. en sélectionnant un sous-ensemble des formes de W_1 et de W_2 qui couvrent toutes les variations dérivationnellement « utiles »

IND.PRS.SG ; IND.PRS.3PL ; IND.IPFV ; PTCP.PRS ; PTCP.PST.M ; PTCP.PST.F

3. en décrivant l'espace thématique de W_1 et de W_2 (= ensembles des radicaux de W_1 et de W_2)

Les transcriptions des formes fléchies sont disponibles dans **GLÀFF** (`http://redac.univ-tlse2.fr/glaffoli/index.jsp`).

Comment décrire les propriétés phonologiques dans Demonette v2 ? (2)

Quel codage phonologique ?

- IPA : facile à lire et produire par les éditeurs, mais il impose une base de données codée en UTF8
- SAMPA : codage de la base de données en ASCII, mais les représentations sont plus complexes à lire et à manipuler ; certains phonèmes sont codés par 2 caractères ;
  mangeait : IPA = $\mathrm{mɑ̃.ʒe}$; SAMPA = `mA~.ZE`

La solution retenue est de fournir une interface d'édition qui effectue la conversion dans les deux sens (IPA $\leftrightarrows$ SAMPA) à la volée.

Comment décrire les propriétés phonologiques dans Demonette v2 ? (3)

Quelle variation radicale ?

- les exposants ne présentent pas de variation
- les variations affectent les radicaux
- il faut identifier le radical dans chaque forme de W_1 et de W_2
- quel niveau d'abstraction ?
  quel codage pour les variations ?