Utilisation de modèles distributionnels pour l'étude du sens des mots morphologiquement construits

Marine Wauquier

Utilisation de modèles distributionnels pour l'étude du sens des mots morphologiquement construits

Marine Wauquier

sous la direction de Nabil Hathout
en collaboration avec Cécile Fabre

CLLE-ERSS, Université de Toulouse & CNRS

4 décembre 2017

Objectifs

- Approche extensive pour la morphologie
  - Morphologie extensive (Plénat et al., 2002; Hathout, 2009)
    - Description linguistique à partir d'un maximum de données
  - Conditionnée par les outils
  - Extensive vis à vis du corpus et non du lexique
- Apports de la sémantique distributionnelle pour l'analyse des procédés dérivationnels et des catégories sémantiques
  - Contributions linguistiques
    - Distinction sémantique des suffixes *-eur*, *-euse* et *-rice*
    - Distinction sémantique des nominalisations en *-age*, *-ion* et *-ment*
  - Contributions méthodologiques

Plan

Plan

L'hypothèse distributionnelle

Harris (1954)

"difference of meaning correlates with difference of distribution"

- Proximité sémantique des mots caractérisée par le partage de contextes
- *parcours* et *itinéraire* sont proches parce qu'ils sont tous les deux
  - sujets des verbes *mener* et *traverser*
  - compléments du nom *étape*
  - modifiés par les adjectifs *spirituel* et *libre*
  - etc.

La sémantique distributionnelle

- Représentation du sens sous la forme de vecteurs
  - À partir des contextes en corpus

	baliser	traverser	changer	étape	incident	spirituel	libre
parcours	0.6	0.3	0.4	0.6	0.7	0.8	0.8
itinéraire	0.7	0.4	0.5	0.6	0.1	0.2	0.2

- Deux mots sont proches si leurs vecteurs sont proches
- Objets mathématiques
  - Mesure de similarité basée sur l'angle entre deux vecteurs
  - Possibilité de les manipuler (addition, soustraction...)

Word2Vec (Mikolov et al., 2013)

- Modèle prédictif
  - Réseau de neurones
    - CBOW : prédiction du mot à partir du contexte
    - Skip-gram : prédiction du contexte
  - Hyper-paramètres
    - Fréquence minimum
    - Nombre de dimensions
  - Mesure de similarité cosinus
    - $0 < P < 1$

Le critère distributionnel pour la recherche en linguistique

- Études en diachronie
  - Évolution du sens (Gulordava and Baroni, 2011; Kulkarni et al., 2015)
  - Évolution de la productivité syntaxique (Perek, 2016)
- Typage sémantique
  - Induction de classes sémantiques verbales (Schulte Im Walde, 2006)
  - Classification de noms composés (Verhoeven et al., 2012)
- Figement sémantique
  - Découpage de composés (Riedl and Biemann, 2016)

Quantification du changement sémantique en diachronie

Kim et al. (2014)

- Le changement du sens est apprécié à partir du traitement d'un corpus diachronique en étudiant le degré de recouvrement des voisins distributionnels des mots calculés par période
- GoogleBooks, 100 milliards de mots

Mot	Voisins en :	
	1900	2009
cell	closet, dungeon, tent	phone, cordless, cellular
gay	cheerful, pleasant, brillant	lesbian, bisexual, lesbians

Du lexème au caractère

Changement d'échelle (Turney and Pantel, 2010)

- Capter du sens à l'échelle des caractères ?

Place de la morphologie

- La morphologie au service de la sémantique distributionnelle
- La sémantique distributionnelle au service de la morphologie

La morphologie au service de la sémantique distributionnelle

- Améliorer les modèles distributionnels en les enrichissant d'informations morphologiques
  - Calcul des vecteurs des mots rares ou absents à partir des vecteurs des lemmes de la famille dérivationnelle (Padó et al., 2013)
  - Représentation des transformations morphologiques sous la forme de vecteurs pour calculer les vecteurs des mots rares (Soricut and Och, 2015)
  - Représentation des séquences de caractères sous la forme de vecteurs et calcul du vecteur d'un mot comme la somme des vecteurs des séquences de caractères (Bojanowski et al., 2016; Luong et al., 2013; Botha and Blunsom, 2014)

Représentation vectorielle des morphèmes

Botha and Blunsom (2014)

Similarité sémantique entre *abstract*, *abstraction* et *abstracted*

- Améliorer la représentation des mots morphologiquement liés
- Représenter les mots absents

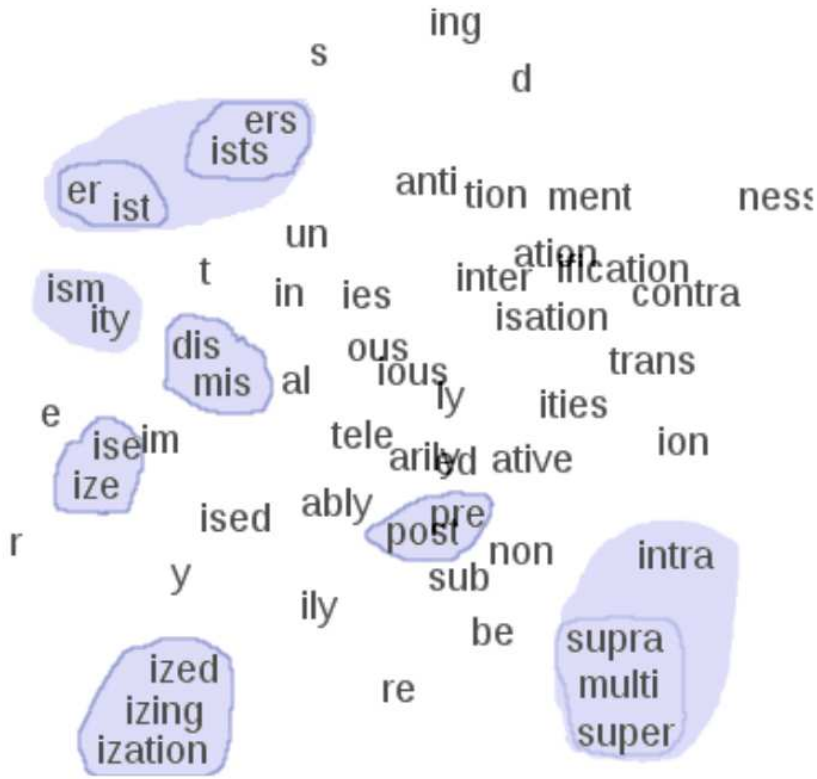
Vecteur du mot comme l'addition des vecteurs de ses composants

- Segmentation morphologique (Morfessor)
- $\overrightarrow{\text{perfectly}} = \overrightarrow{\textit{perfectly}} + \overrightarrow{\textit{perfect}} + \overrightarrow{\textit{ly}}$

Représentation vectorielle des morphèmes

Botha and Blunsom (2014)

- Visualisation des vecteurs d'affixes après réduction de dimensions



La sémantique distributionnelle au service de la morphologie

- Étudier par le biais des modèles distributionnels des questions en morphologie
  - Validation sémantique de ressources dérivationnelles (Zeller et al., 2014)
  - Déterminer la direction de procédés dérivationnels (Kisselew et al., 2016; Padó et al., 2015)
  - Comparaison sémantique de procédés dérivationnels (Kisselew et al., 2015; Varvara et al., 2016; Botha and Blunsom, 2014)
  - Désambiguïsation des interprétations (Lapesa et al., 2017)

Comparaison de procédés dérivationnels

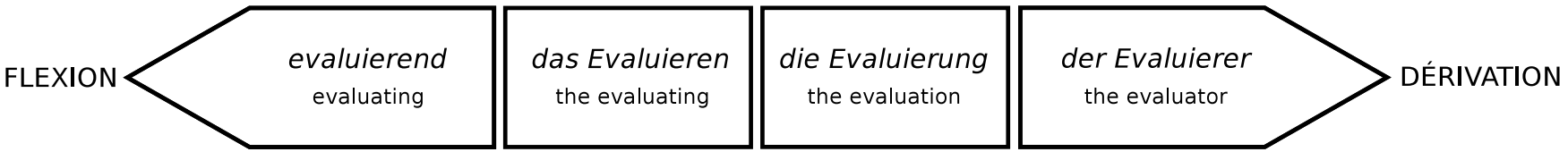
Varvara et al. (2016)

Comparaison sémantique de procédés de nominalisation similaires

- Infinitif nominal (NI) : *das Evaluieren* (*the evaluating*)
- Déverbal en -*ung* (UNG) : *die Evaluierung* (*the evaluation*)

Positionnement de ces procédés vis à vis de la flexion et de la dérivation

- Participe présent (PP) : *evaluierend* (*evaluating*)
- Nom d'agent en -*er* (ER) : *der Evaluierer* (*the evaluator*)

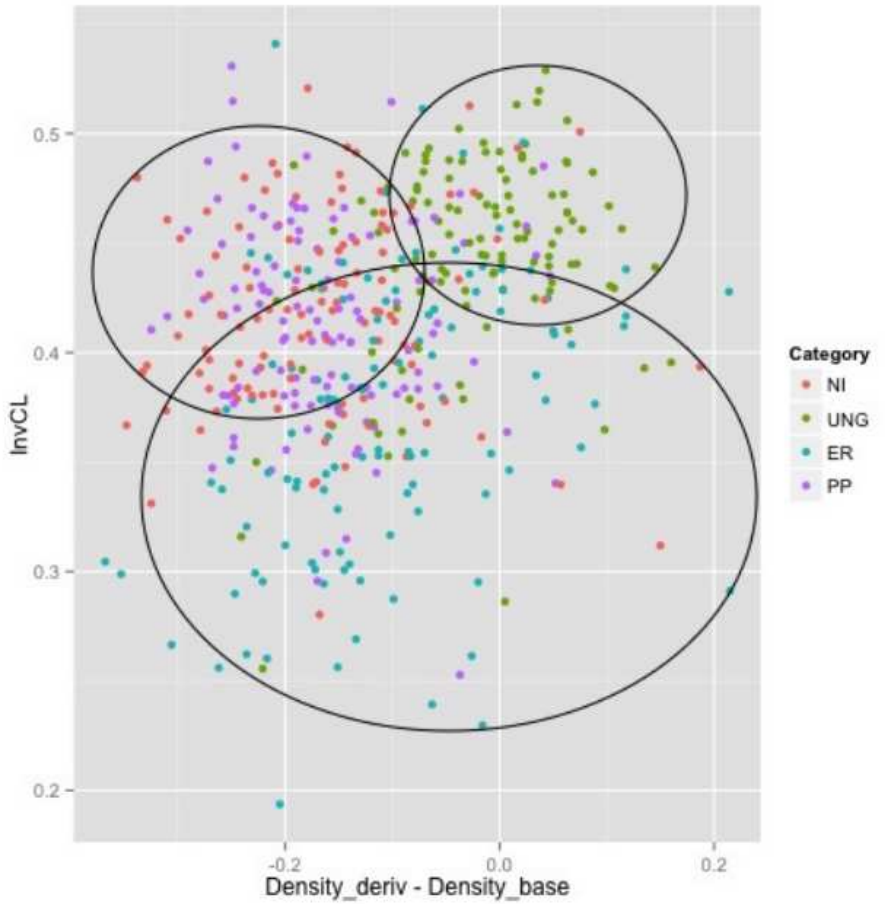


Varvara et al. (2016)

- SDeWaC lemmatisé : 846 millions de mots
- Quadruplets NI, UNG, PP et ER extraits de DerivBase (Zeller et al., 2013)
  - Filtrage par la fréquence (2 écart-types de la fréquence moyenne)
  - 115 quadruplets
- Modèle "*count-based*"
  - Transparence : *feature inclusion*
  - Spécificité : densité du voisinage

Comparaison de procédés dérivationnels

Varvara et al. (2016)



Désambiguïsation en contexte des nominalisations en *-ment*

Lapesa et al. (2017)

Distinguer les néologismes en *-ment* événementiels et non événementiels

- *In many places,* ***emplacement*** *of granite plutons is synchronous to volcanic eruptions.* (EVENTIVE)
- *I set down the scrap of dolls dress, a* ***bedragglement*** *of loose lace hem* (NON EVENTIVE)

Désambiguïsation en contexte des nominalisations en *-ment*

Lapesa et al. (2017)

- Extraction de candidats en *-ment*
  - COCA et OED
  - Filtrage par la catégorie du verbe
    - Verbes d'états psychologiques (*admire*), impliquant de la force (*push*), de changement d'état (*break*) et *putting verbs* (*put*)
  - Extraction d'occurrences dans GloWbE (1.9 milliard de mots), WebCorp et Google
  - 55 lexèmes et 406 tokens
- Annotation manuelle de chaque occurrence

Lapesa et al. (2017)

- Calcul des vecteurs d'occurrence
  - Moyenne des vecteurs de contexte
  - "*The suit is next to the tie and the t-shirt.*" vs "*The lawyer filed a suit to the judge.*"
- Entraînement d'un classifieur à partir des vecteurs de noms d'événement et d'entités typiques
  - BNC et ukWaC, 2.6 millliards de mots
- Prédiction de la classe des vecteurs d'occurrences des dérivés en *-ment*
  - *Hydrogen, especially atomic hydrogen, is particularly dangerous because it tends to cause rapid* ***embrittlement*** *even at low temperatures. (0.96)*

Plan

Proximité sémantique de dérivés morphologiques

Roché (2009)

Le verbe est sémantiquement plus proche du nom d'action que du nom d'agent

- *protéger* et *protection* plus proches que *protéger* et *protecteur*
  - *avec pour mission de* ***protéger*** *Wyatt*
  - *demandent que la France assure la* ***protection*** *de Ayaan Hirsi Ali*
  - *Puis il devient le* ***protecteur*** *de Raf (Raphaël) Esquivel*

Quantifier la proximité sémantique des mots deux par deux au sein du triplet verbe-action-agent

- P(Verbe-Action) > P(Verbe-Agent)
  - P(*protéger-protection*) > P(*protéger- protecteur*) ?
  - P(*nager-nage*) > P(*nager- nageur*) ?
  - P(*confectionner-confection*) > P(*confectionner- confectionneur*) ?
  - ...
- P(Verbe-Action) > P(Action-Agent)
  - P(*protéger-protection*) > P(*protection- protecteur*) ?
  - P(*nager-nage*) > P(*nage- nageur*) ?
  - P(*confectionner-confection*) > P(*confection- confectionneur*) ?
  - ...

Lexeur (1)

- Sélection des noms d'agents déverbaux en *-eur* grâce à Lexeur

Nom d'agent masc.	Nom d'agent fém.	Base	Cat.	Autres dérivés
abatteur/Ncms	abatteuse/Ncfs	abattre/Vmn--	Vb	abat/Ncms ; abattement/Ncms ; abatture/Ncfs ; abattage/Ncms ; abattis/Ncms
endoscopeur/Ncms	endoscopeuse/Ncfs	∅	∅	endoscopie/Ncfs
fraudeur/Ncms	fraudeuse/Ncfs	frauder/Vmn--	Vb	fraude/Ncfs
sculpteur/Ncms	sculpteuse/Ncfs ; sculptrice/Ncfs	sculpter/Vmn--	Vb	sculpture/Ncfs ; sculptage/Ncms
wheeleur/Ncms	wheeleuse/Ncfs	wheel/Ncms	Nb	∅

Figure : Extrait de Lexeur

Lexeur (2)

- Couverture
  - 5974 entrées : noms d'agent en *-eur*
    - Base : 79% de bases verbales
    - Féminin : 75% en *-euse* vs 25% en *-rice*
    - Action : 79% sont associées à au moins un nom d'action (entre 1 et 8)
- Qualité
  - Constitution manuelle à partir de TLFi et d'attestations issues du Web
  - Par des linguistes
- Limites
  - Erreurs liées à sa constitution
    - Familles dérivationnelles : *ouvreur*, *ouvreuse*, *ouvrir*, *ouverture*, *ouvrage*
    - Doublons : *énumérateur*
    - Annotation : *collision* annotée *Xb* au lieu de *Nb*
  - Polysémie
    - *soudeuse*, *traceur*, *construction*

Word embeddings

- Word2Vec (Mikolov et al., 2013)
  - Contexte graphique
  - CBOW
  - Fréquence minimum de 5
  - 100 dimensions
- Corpus
  - Lemmatisés
  - *Wikipedia*
    - Version fr de l'encyclopédie en ligne
    - 255 millions de mots
  - *LM10*
    - Articles publiés entre 1991 et 2000 du journal *Le Monde*
    - 200 millions de mots

Démarche

- Extraction des triplets Agent-Verbe-Action (13 136)
- Filtrage des triplets
  - 1945 triplets présents dans le modèle *Wikipedia*
  - 1520 triplets présents dans le modèle *LM10*
- Calcul des scores de proximité P
- Attribution d'une étiquette sur la base du score P le plus élevé

Agent	Verbe	Action	P(AgVb)	P(AgAc)	P(VbAc)	Tag
discriminateur	discriminer	discrimination	0.15	0.15	0.49	VbAc
directeur	diriger	direction	0.44	0.52	0.27	AgAc

Table : Exemple de triplets étiquetés

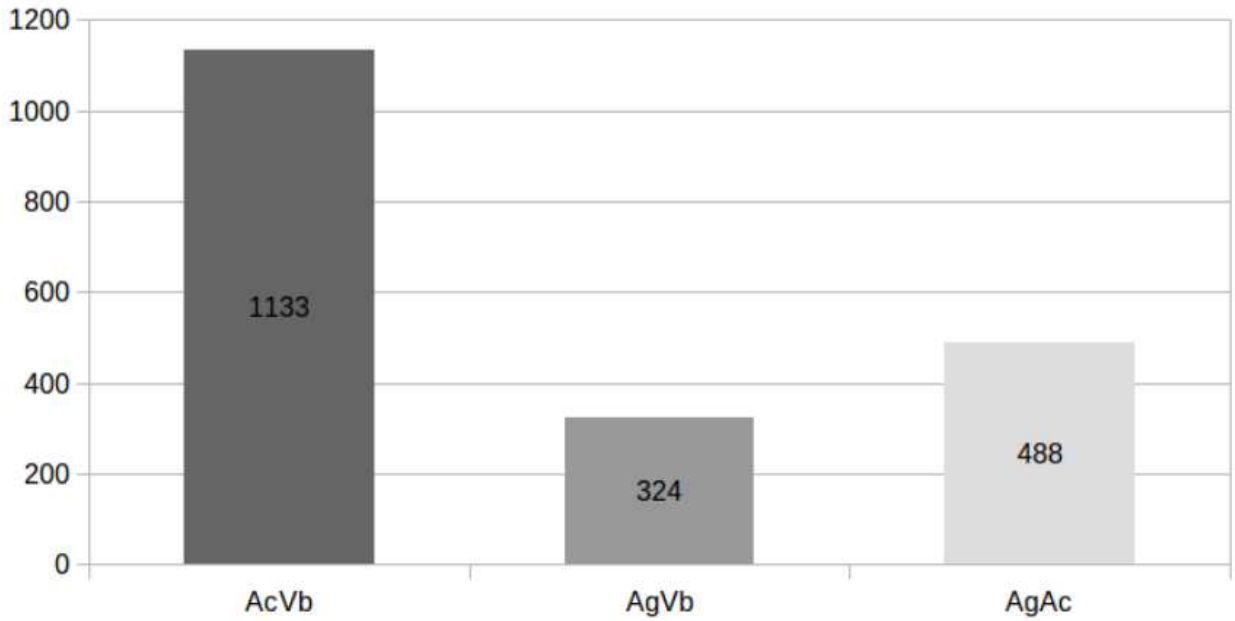
Démarche

- Extraction des triplets Agent-Verbe-Action (13 136)
- Filtrage des triplets
  - 1945 triplets présents dans le modèle *Wikipedia*
  - 1520 triplets présents dans le modèle *LM10*
- Calcul des scores de proximité P
- Attribution d'une étiquette sur la base du score P le plus élevé

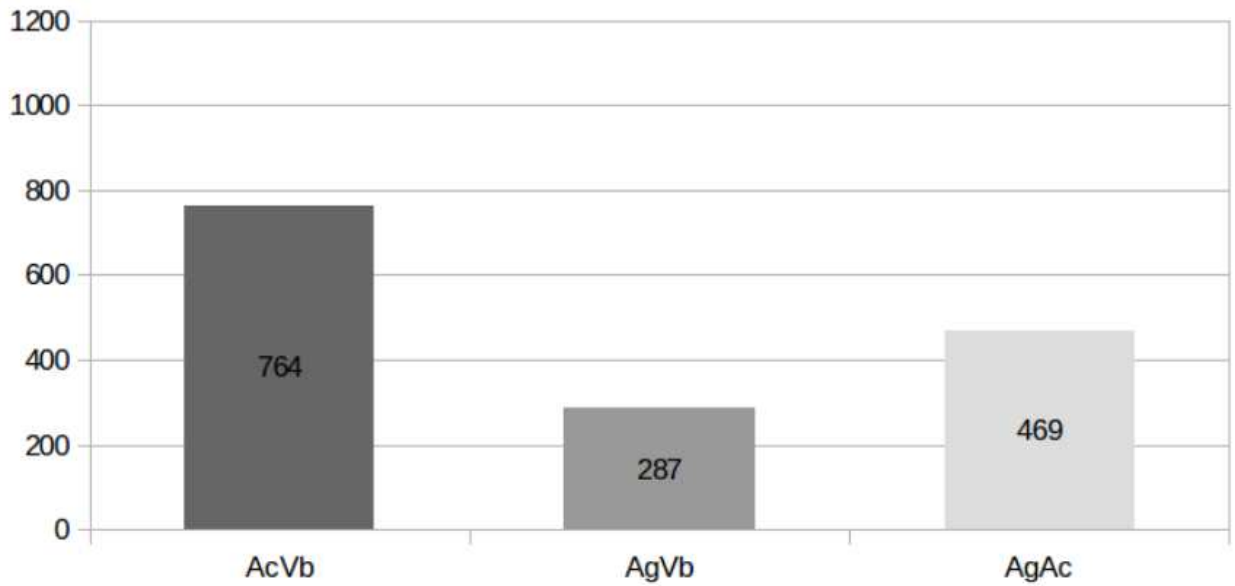
Agent	Verbe	Action	P(AgVb)	P(AgAc)	P(VbAc)	Tag
discriminateur	discriminer	discrimination	0.15	0.15	0.49	VbAc
directeur	diriger	direction	0.44	0.52	0.27	AgAc

Table : Exemple de triplets étiquetés

Répartition des triplets issus de *Wikipedia*



Répartition des triplets issus de *LM10*



Indices de proximité moyens

	P(AgVb)	P(AgAc)	P(VbAc)
Wikipedia	0.25	0.29	**0.39**
LM10	0.25	0.28	**0.34**

Exemples

- Cas conformes à nos attentes

Verbe	Action	Agent	P(AgVb)	P(AgAc)	P(VbAc)
protéger	protection	protecteur	0.45	0.37	0.59
nager	nage	nageur	0.45	0.54	0.73
confectionner	confection	confectionneur	0.42	0.54	0.62

Exemples

- Impact de la fréquence

Verbe	Action	Agent	P(AgVb)	P(AgAc)	P(VbAc)
concevoir	conception	conceptrice	0.10	0.03	0.49

Exemples

- Impact de la fréquence

Verbe	Action	Agent	P(AgVb)	P(AgAc)	P(VbAc)
concevoir	conception	conceptrice	0.10	0.03	0.49
concevoir	conception	concepteur	0.60	0.51	0.49

Exemples

- Impact de la fréquence

Verbe	Action	Agent	P(AgVb)	P(AgAc)	P(VbAc)
concevoir	conception	conceptrice	0.10	0.03	0.49
concevoir	conception	concepteur	0.60	0.51	0.49

concevoir	conception	concepteur	conceptrice
20998	15503	2247	7

Exemples

- Impact de la polysémie et du POS

Verbe	Action	Agent	P(AgVb)	P(AgAc)	P(VbAc)
raffiner	raffinement	raffineur	0.23	0.06	0.64

- En corpus
  - *l'électrolyse sert à* ***raffiner*** *l'aluminium*
  - *penser qu'une réalisation d'un tel* ***raffinement*** *et d'un tel gigantisme*
  - *famille de neuf enfants dont le père est* ***raffineur*** *de sucre*

Aide à la révision de la ressource

- Triplets non valides

Agent	Verbe	Action	P(AgVb)	P(AgAc)	P(VbAc)
ouvreur	ouvrir	ouvrage	0.04	0.04	0.01
accepteur	accepter	acception	<0.01	0.06	0.01

- Triplets valides mais...

Agent	Verbe	Action	P(AgVb)	P(AgAc)	P(VbAc)
menteur	mentir	menterie	0.42	0.21	0.01
directeur	diriger	direction	0.44	0.52	0.27

Conclusions de l'expérience

- Confirmation d'une hypothèse linguistique
  - Facilité de mise en place
  - Approche quantitative
  - Applicable à la morphologie
- Importance des ressources
  - Absence d'accès au contexte
  - Valeur de l'interprétation par le biais des voisins

Plan

Peut-on représenter une série dérivationnelle ?

- Calcul de proximité sémantique à l'échelle d'une classe de mots (agent, verbe, action)
- Calcul de proximité sémantique à l'échelle d'une série dérivationnelle

	Verbe	***-age***	***-ion***	***-ment***
Verbe		0.354	0.445	0.372
-eur	0.271	0.313	0.348	0.253
-euse	0.180	0.229	0.186	0.142
-rice	0.187	0.307	0.239	0.175

- Comment représenter le sens prototypique des dérivés et suffixes ?

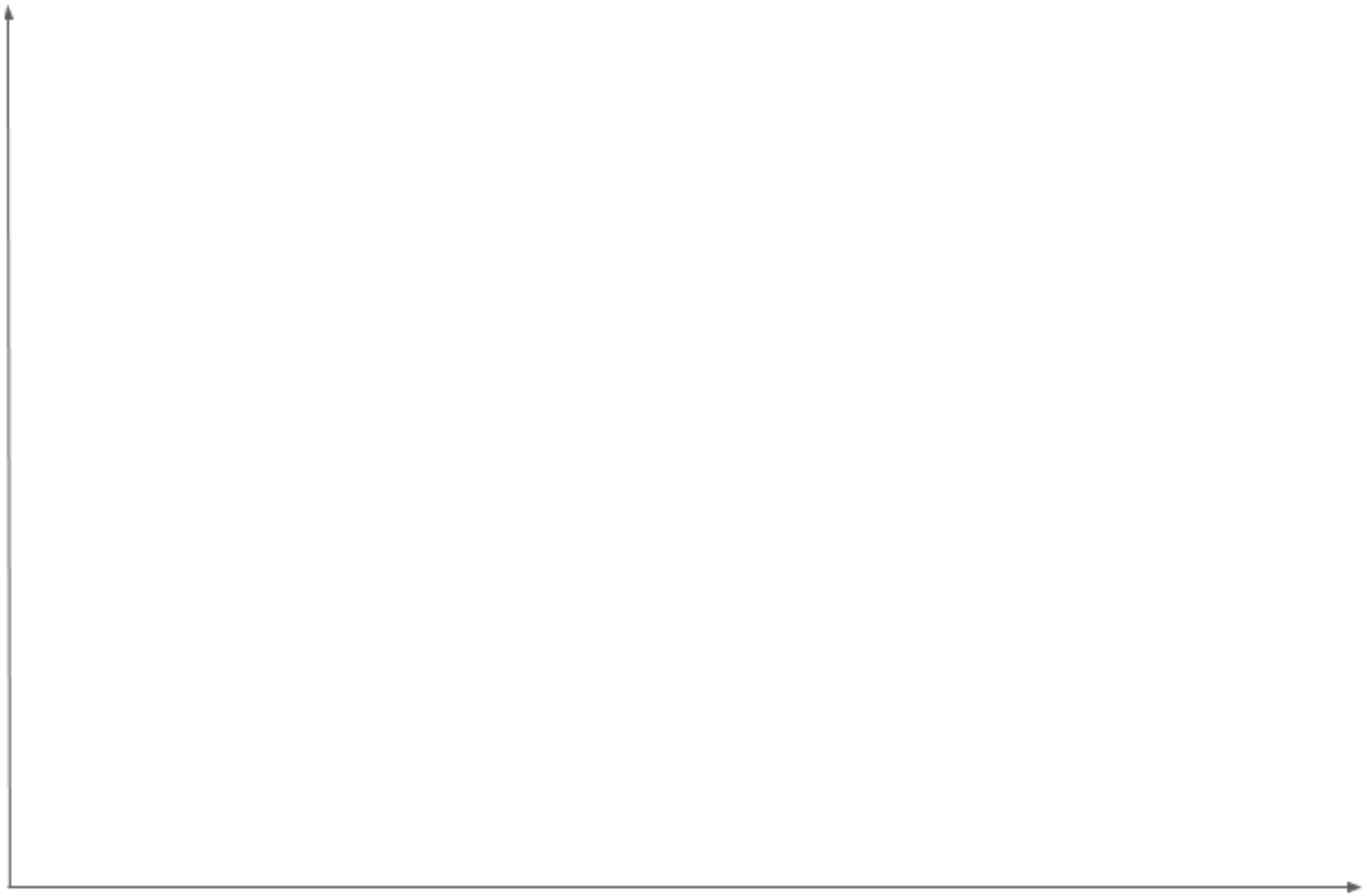
Sens prototypique du dérivé

- Représenter l'identité sémantique prototypique de la classe des noms d'agent en *-eur*, *-euse* et *-rice* en passant par l'analyse des voisins les plus proches
- Une représentation abstraite pour une entité abstraite : cœur d'une classe sémantique
  - Construire un vecteur à partir de word embeddings
  - Vecteur moyen pour représenter le sens "moyen" d'un groupe de mots (Kintsch, 2001)

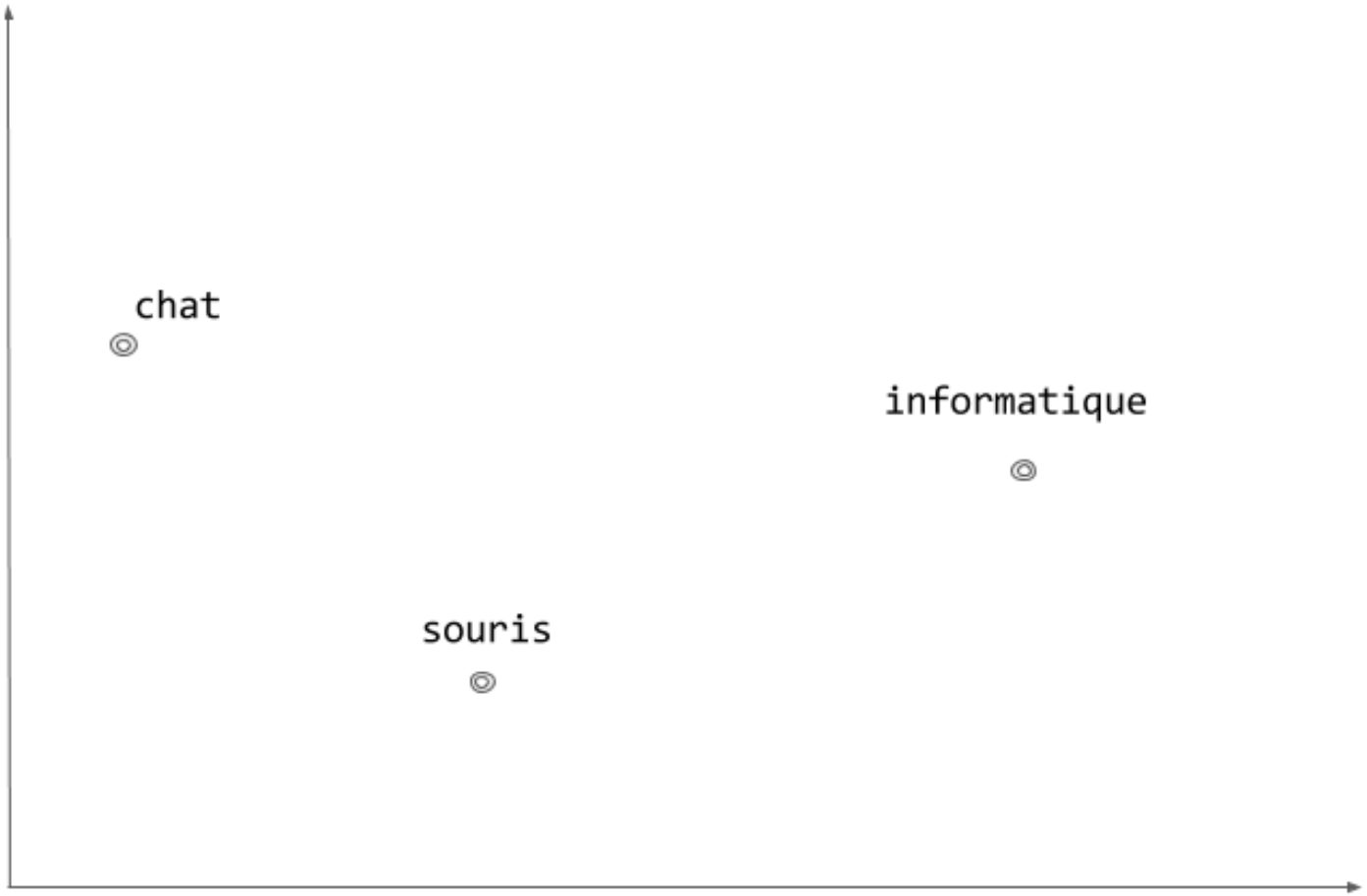
$$\overrightarrow{SUFF} = \frac{\overrightarrow{\text{Nsuff}_1} + \overrightarrow{\text{Nsuff}_2} + ... + \overrightarrow{\text{Nsuff}_n}}{n}$$

- Accès au sens grâce aux voisins les plus proches

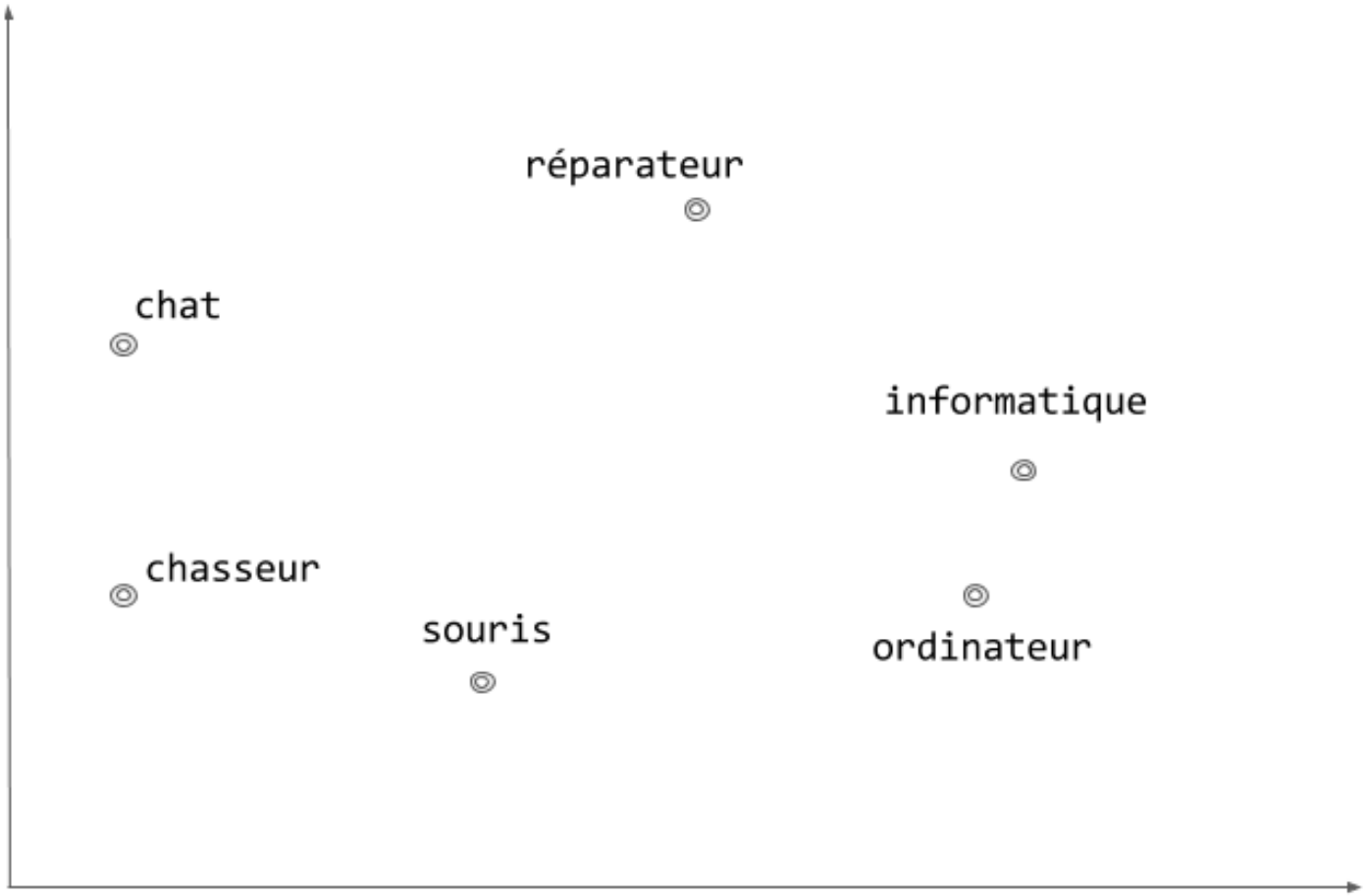
Illustration



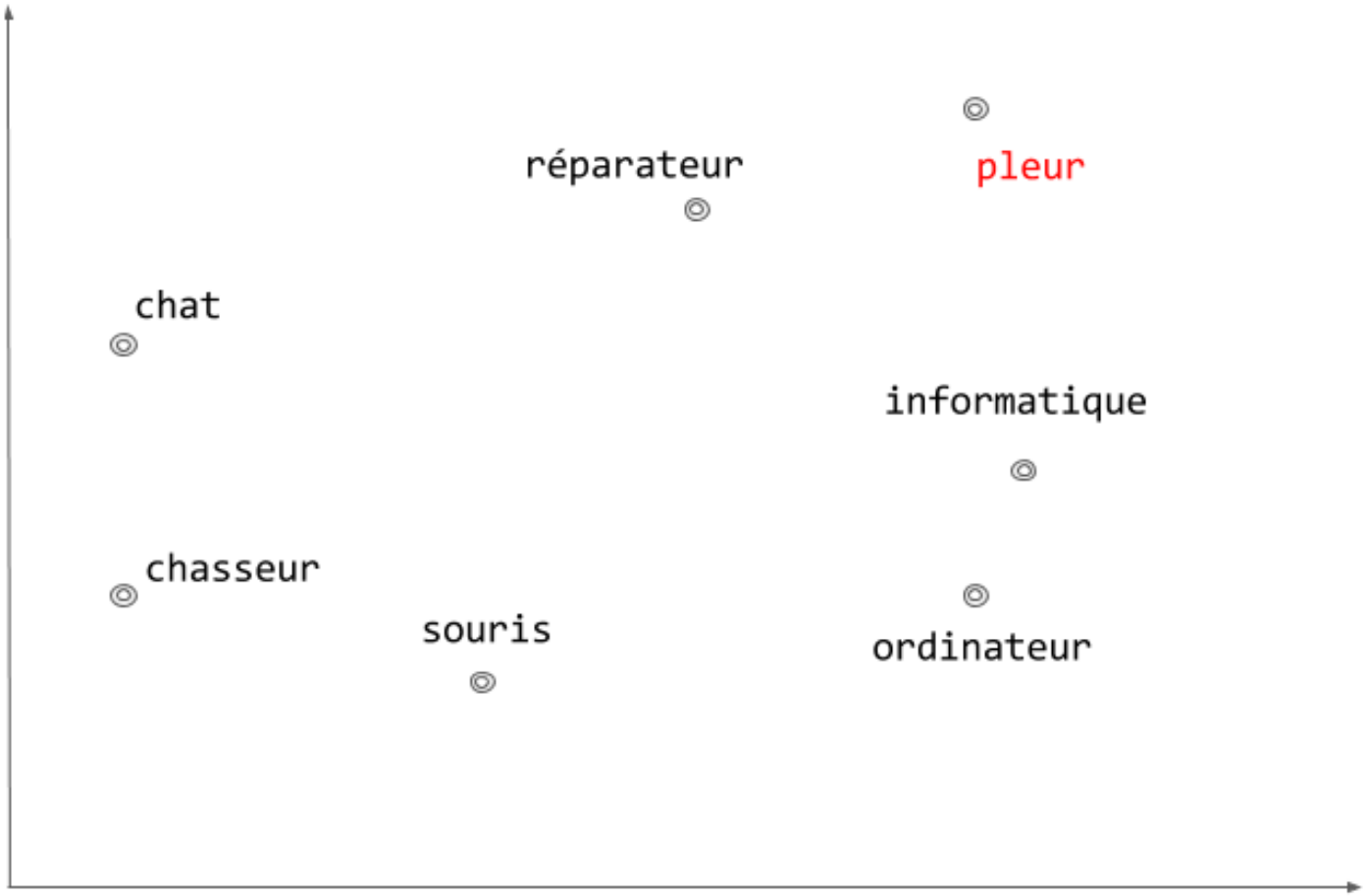
Illustration



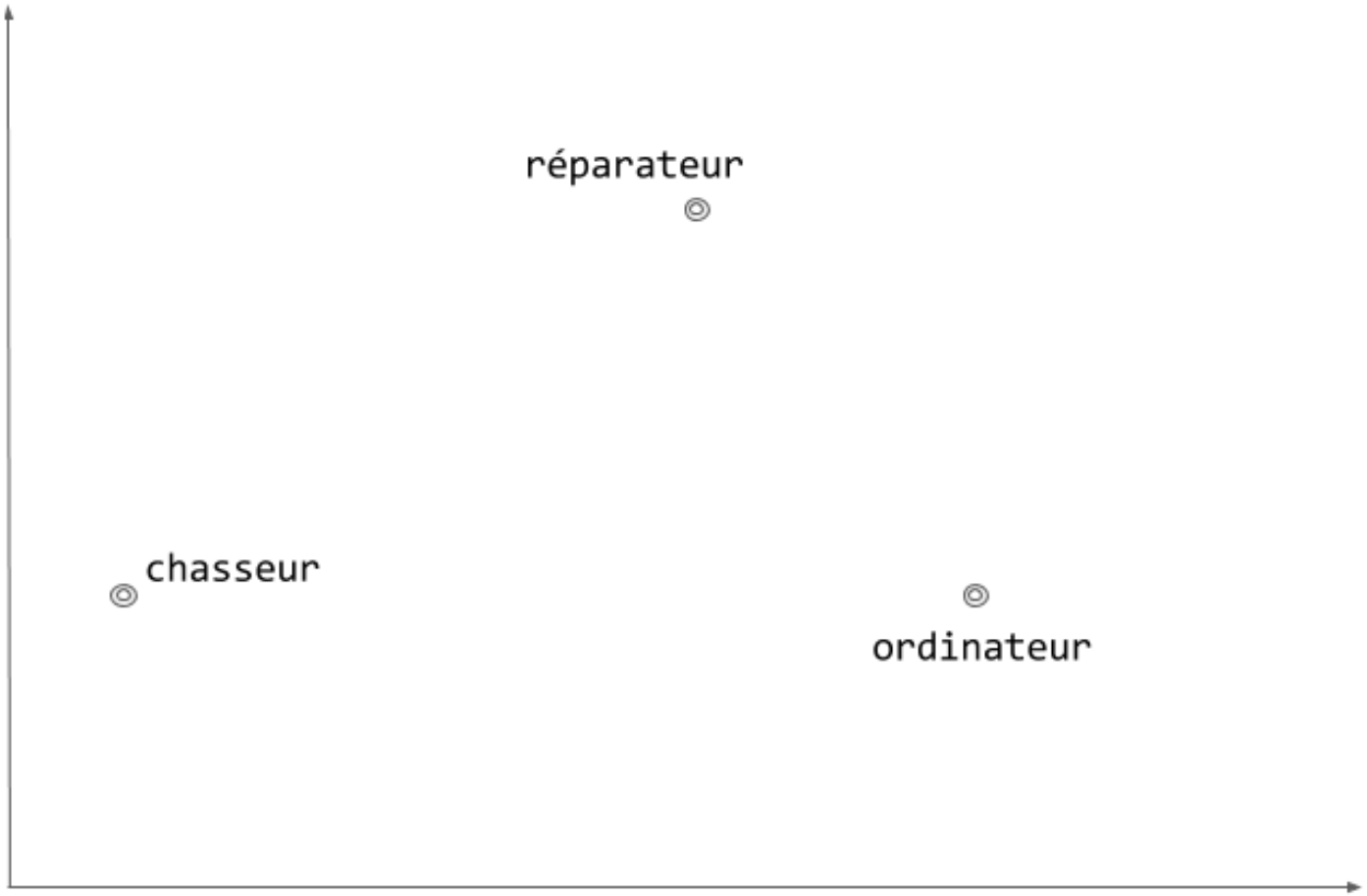
Illustration



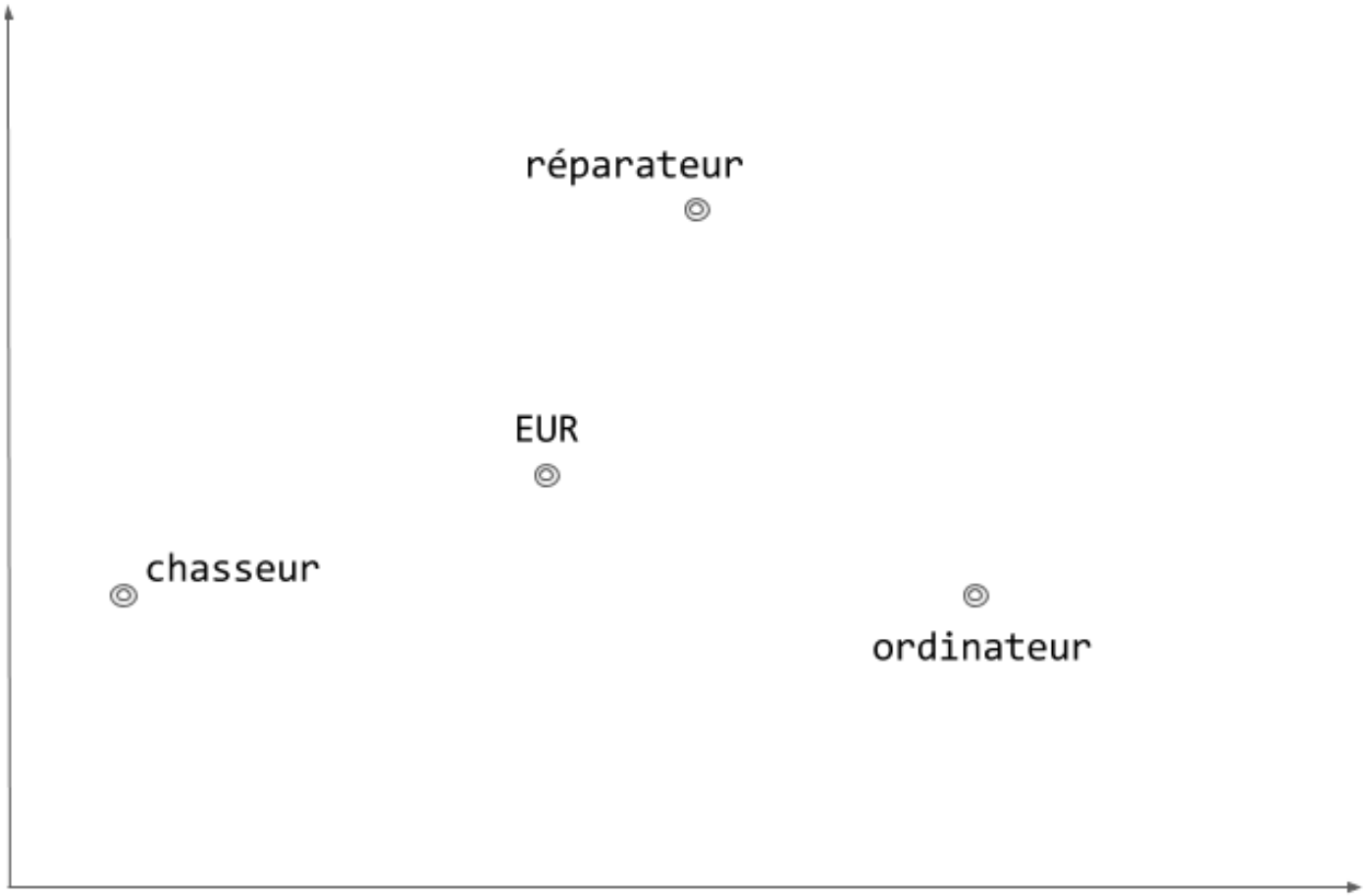
Illustration



Illustration

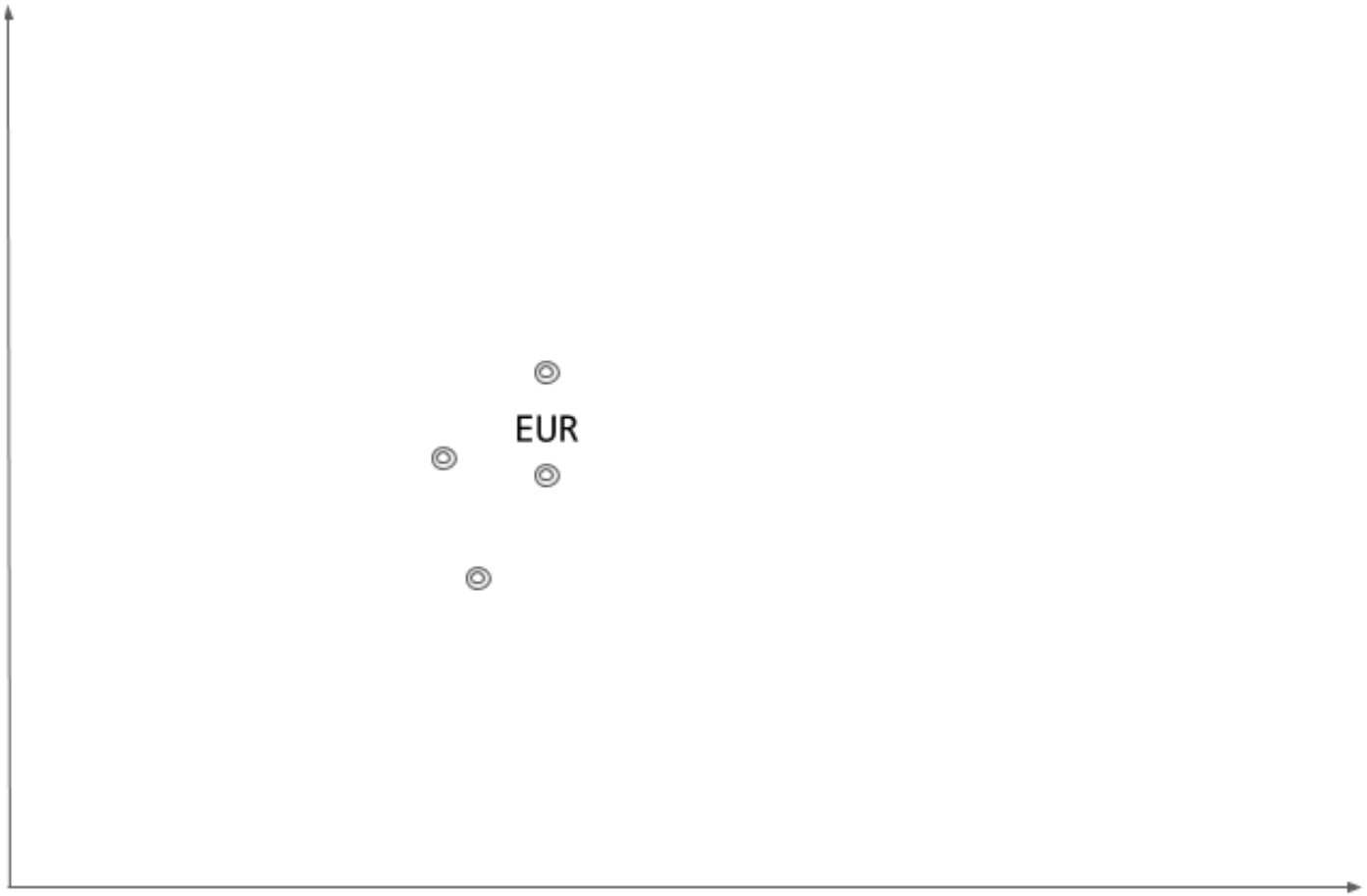


Illustration



Illustration

Illustration



Sens prototypique du suffixe

Représenter l'instruction sémantique du suffixe

- Dérivation : conservation ou ajout de sens (Laca, 2001; Koontz-Garboden, 2007)

$$\text{nominalisation} = \text{verbe} + \text{affixe}$$

- Application par une méthode soustractive (Bolukbasi et al., 2016)

$$\overrightarrow{suff} = \frac{(\overrightarrow{\text{Nsuff}_1} - \overrightarrow{\text{V}_1}) + (\overrightarrow{\text{Nsuff}_n} - \overrightarrow{\text{V}_n}) + ... + (\overrightarrow{\text{Nsuff}_n} - \overrightarrow{\text{V}_n})}{n}$$

Plan

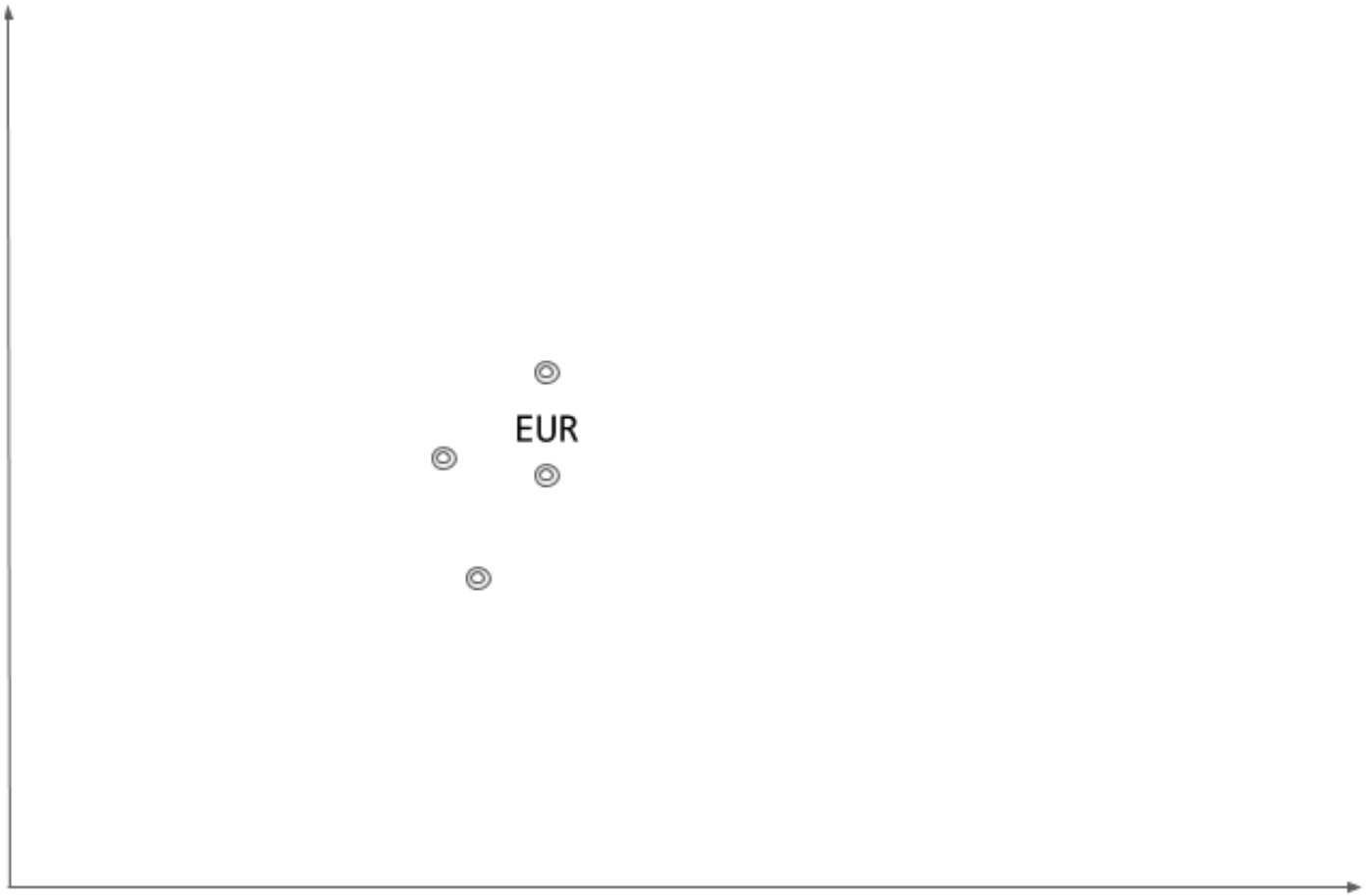
Suffixe *-eur*

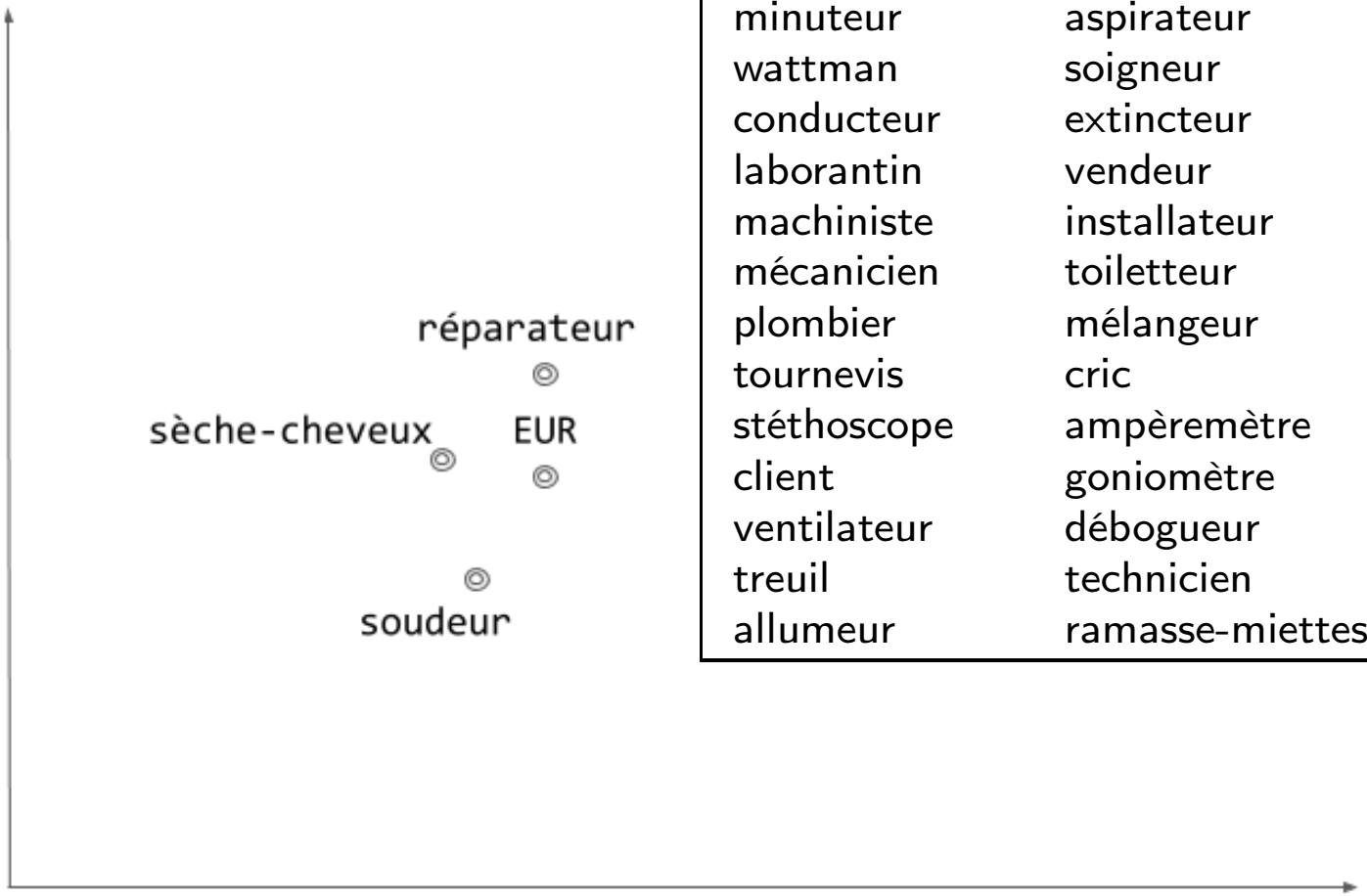
- Suffixation en *-eur*
  - Base verbale ou nominale
    - Thème de présent (*sauveur* => *qui sauve*)
    - Thème de supin (*sauveteur* => *qui fait des sauvetages*) (Benveniste, 1975)
  - Désigne un agent (*traducteur*) ou un instrument (*transmetteur*) (Huyghe and Tribout, 2015)

La féminisation délicate du suffixe *-eur*

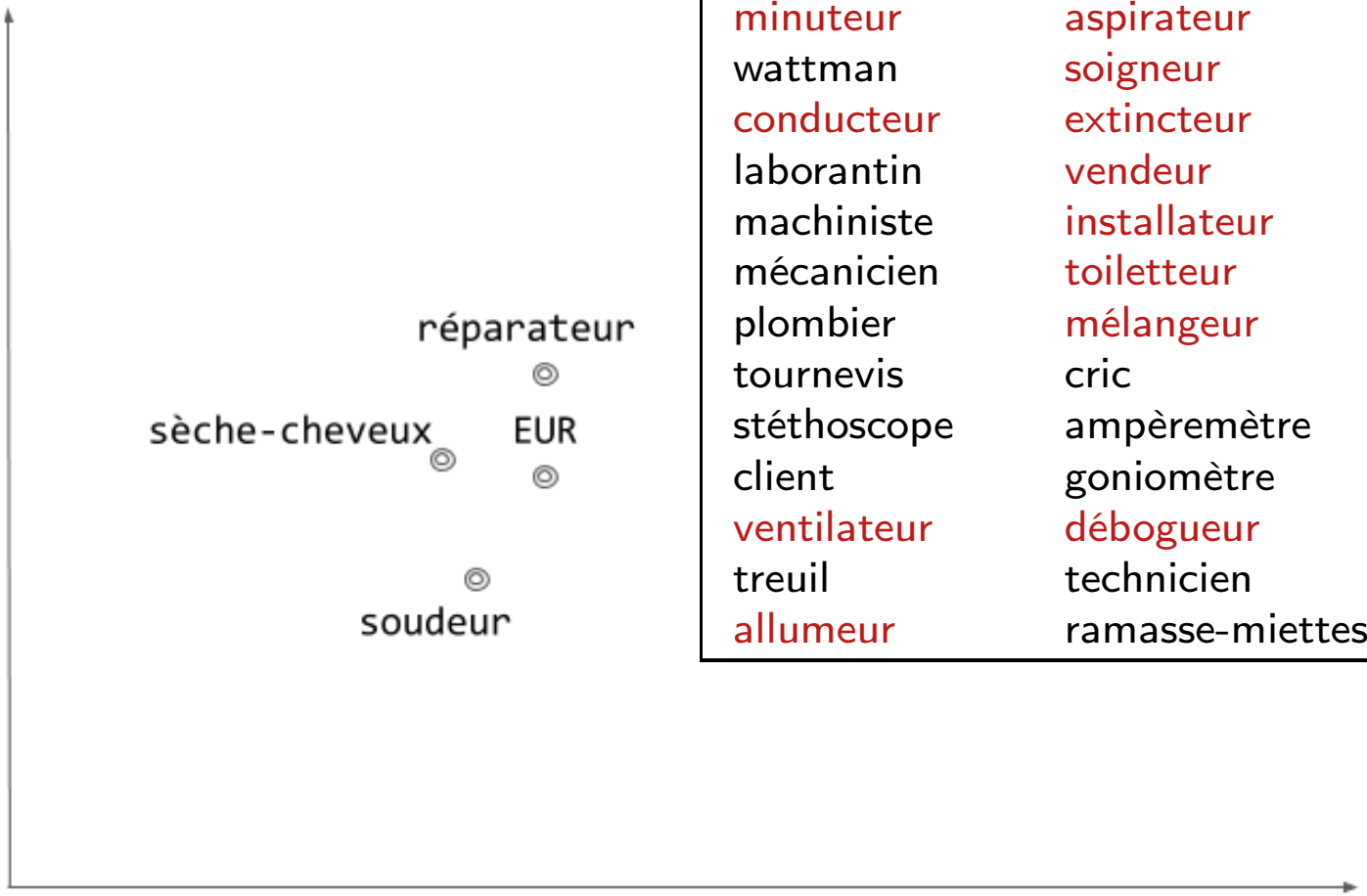
- Une glose
  - Marquage du genre sexuel (Mel'čuk, 2000)
  - Désignait historiquement les instruments (*moissonneuse*) ou la femme de l'agent (*ambassadrice*) (Dubois, 1962; Le Draoulec and Péry-Woodley, 2016)
- Deux suffixes
  - Thème savant pour *-rice* et thème populaire pour *-euse*
  - Connoté (Dawes, 2003; Lenoble-Pinson, 2008)
    - ***-euse*** **:** Dépréciatif, métiers de basse condition (*repasseuse*)
    - ***-rice*** **:** Plus noble, métiers valorisants ou valorisés (*directrice*)

	-eur	***-euse***	***-rice***
Wikipedia	1334	239	90
LM10	1147	155	65

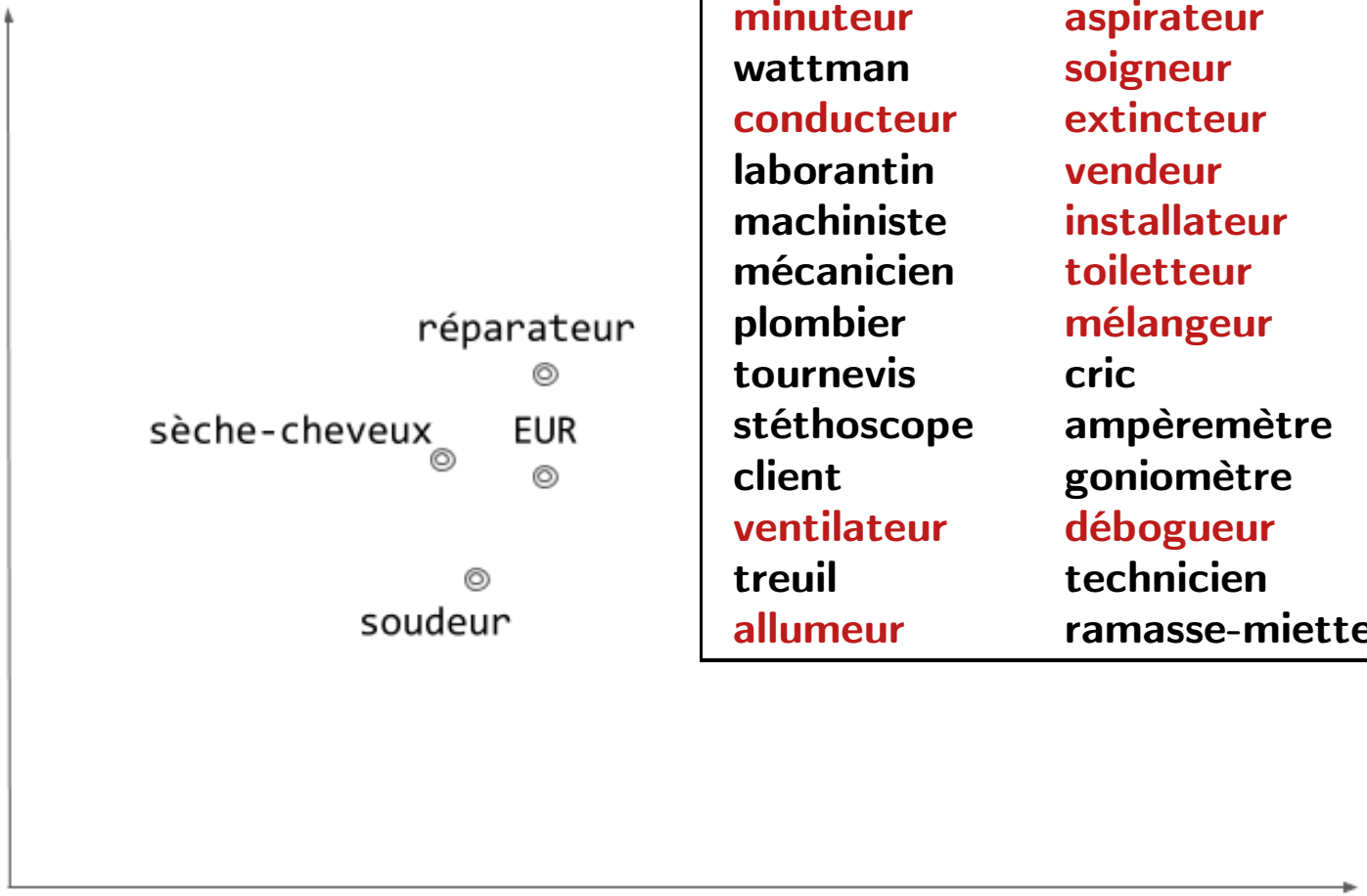




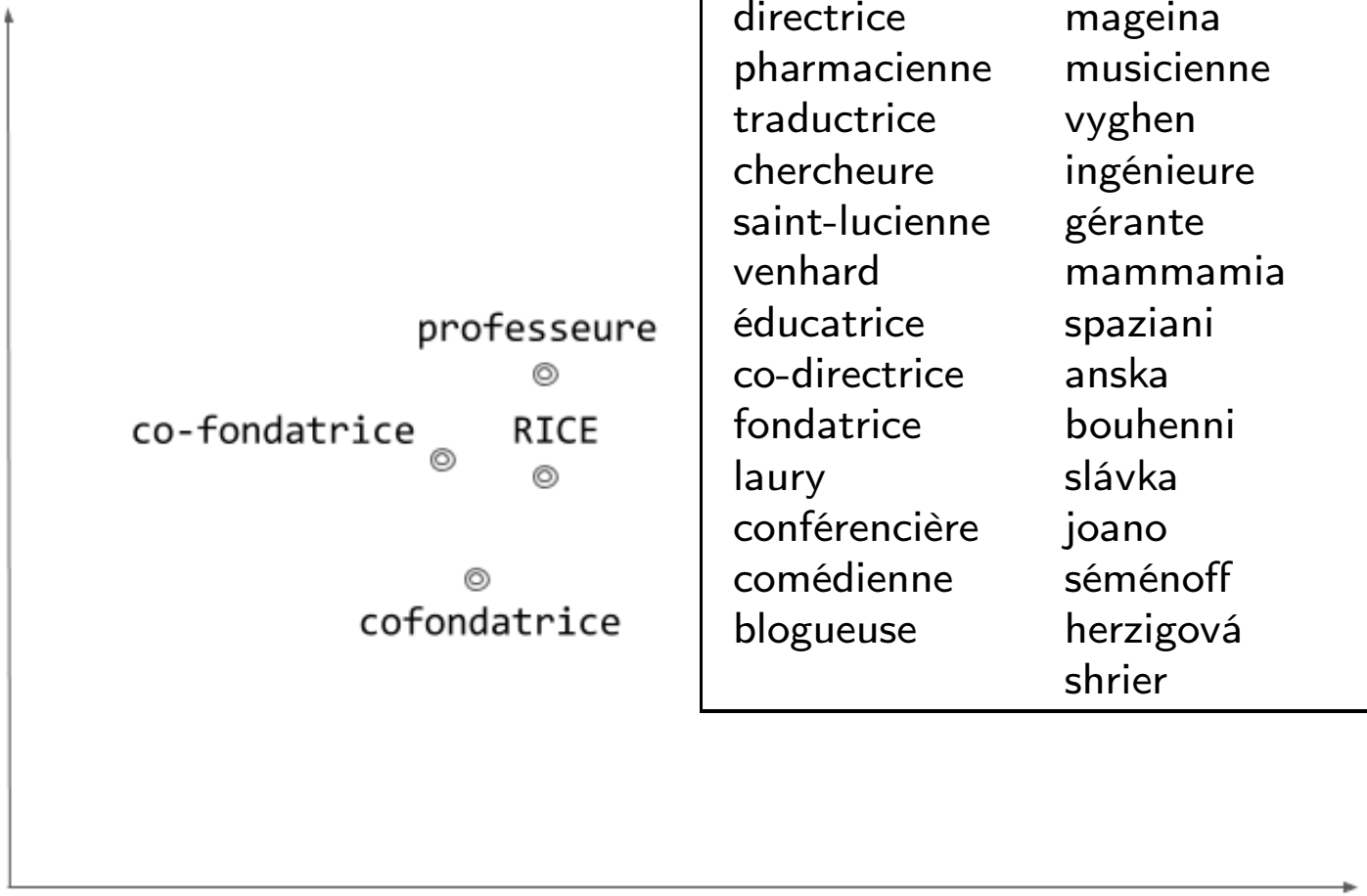
réparateur	mécano	contacteur
sèche-cheveux	coursier	descendeur
soudeur	déménageur	dépresseur
armurier	manomètre	tune-o-matic
minuteur	aspirateur	leurre
wattman	soigneur	télérupteur
conducteur	extincteur	coupe-ongles
laborantin	vendeur	égoutier
machiniste	installateur	microphone
mécanicien	toiletteur	juge-arbitre
plombier	mélangeur	opticien
tournevis	cric	nettoyeur
stéthoscope	ampèremètre	adaptateur
client	goniomètre	grappin
ventilateur	débogueur	détecteur
treuil	technicien	ordinateur
allumeur	ramasse-miettes	



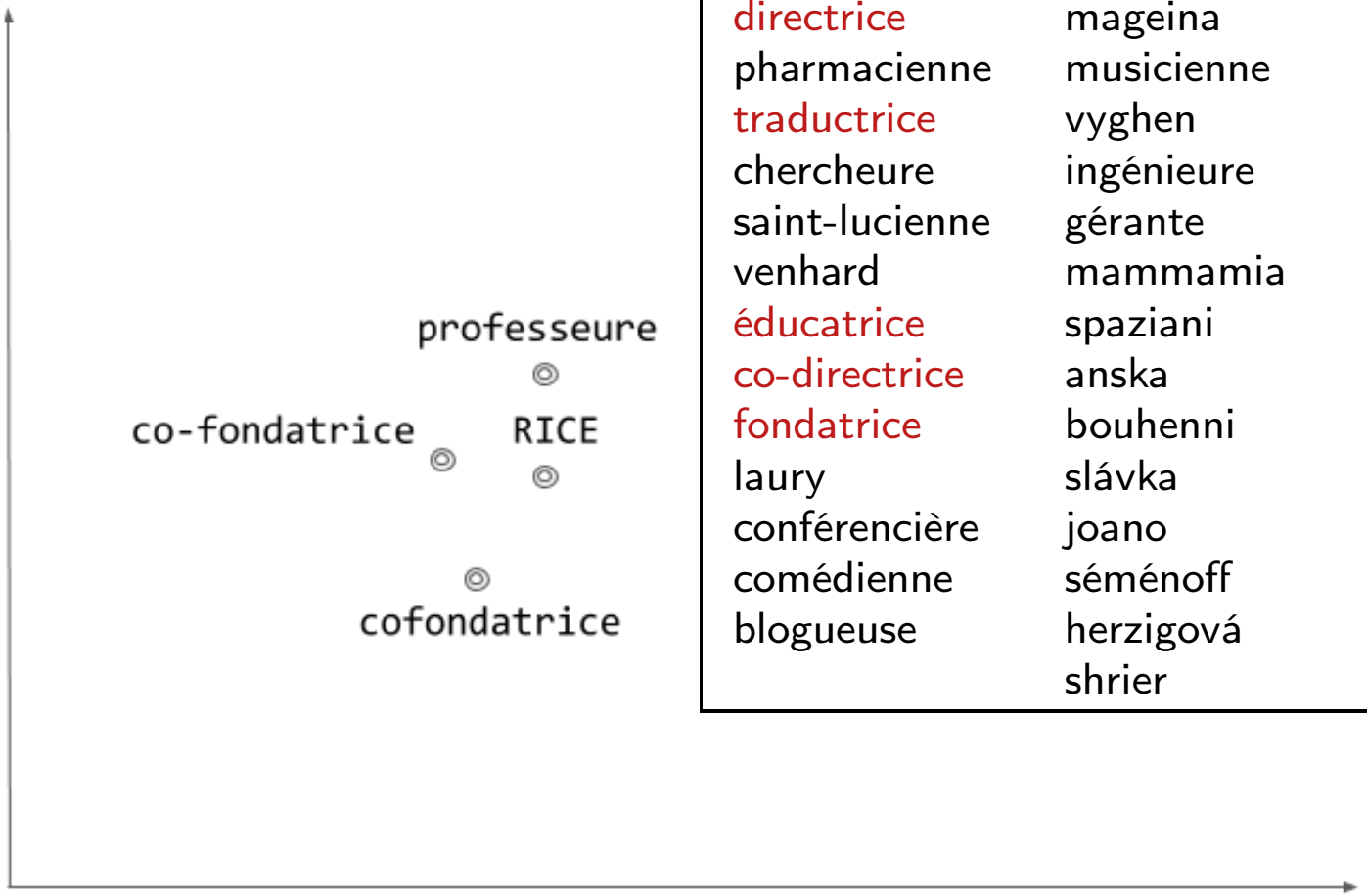
réparateur	mécano	contacteur
sèche-cheveux	coursier	descendeur
soudeur	déménageur	dépresseur
armurier	manomètre	tune-o-matic
minuteur	aspirateur	leurre
wattman	soigneur	télérupteur
conducteur	extincteur	coupe-ongles
laborantin	vendeur	égoutier
machiniste	installateur	microphone
mécanicien	toiletteur	juge-arbitre
plombier	mélangeur	opticien
tournevis	cric	nettoyeur
stéthoscope	ampèremètre	adaptateur
client	goniomètre	grappin
ventilateur	débogueur	détecteur
treuil	technicien	ordinateur
allumeur	ramasse-miettes	



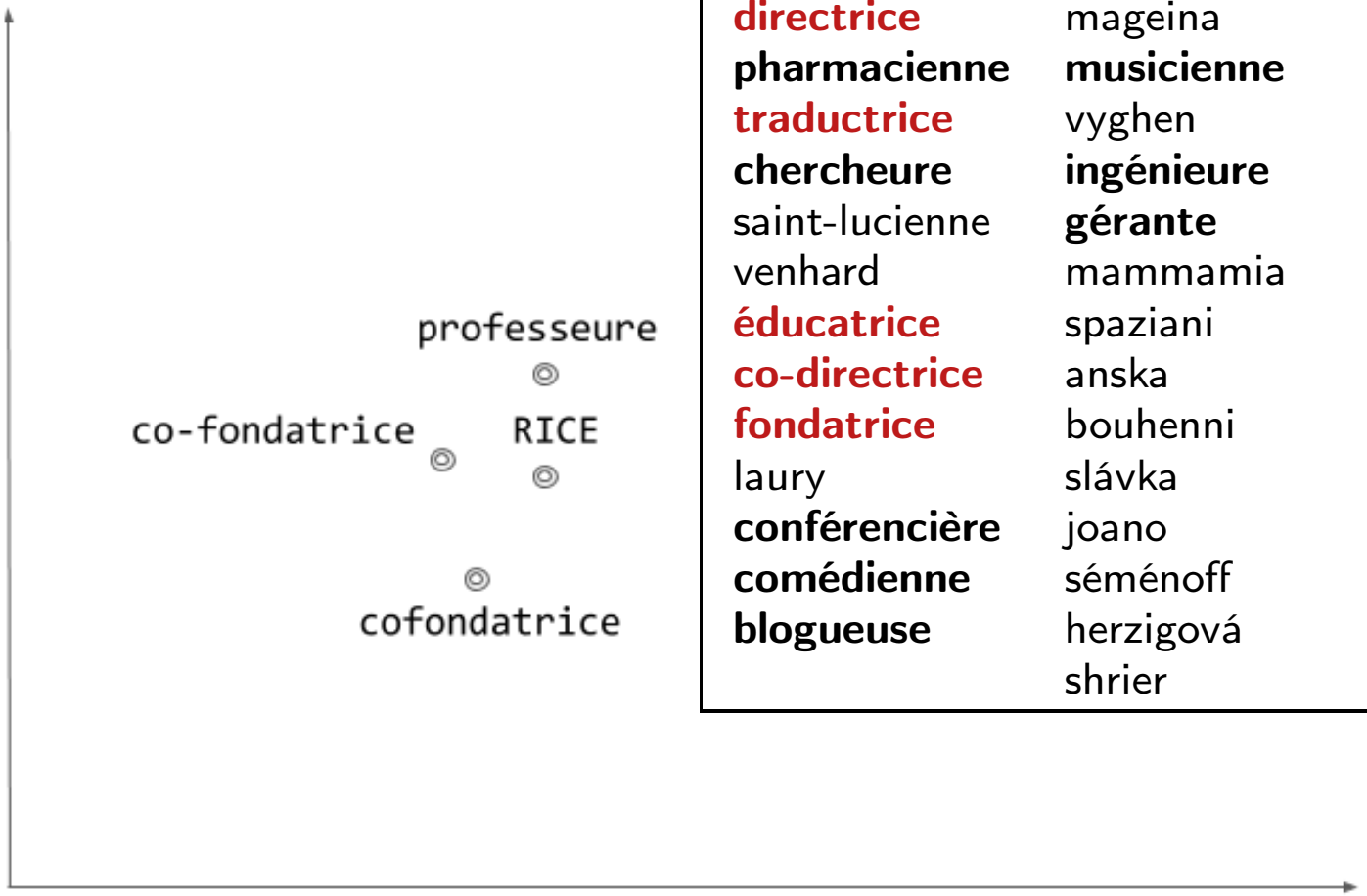
réparateur	mécano	contacteur
sèche-cheveux	coursier	descendeur
soudeur	déménageur	dépresseur
armurier	manomètre	tune-o-matic
minuteur	aspirateur	leurre
wattman	soigneur	télérupteur
conducteur	extincteur	coupe-ongles
laborantin	vendeur	égoutier
machiniste	installateur	microphone
mécanicien	toiletteur	juge-arbitre
plombier	mélangeur	opticien
tournevis	cric	nettoyeur
stéthoscope	ampèremètre	adaptateur
client	goniomètre	grappin
ventilateur	débogueur	détecteur
treuil	technicien	ordinateur
allumeur	ramasse-miettes	



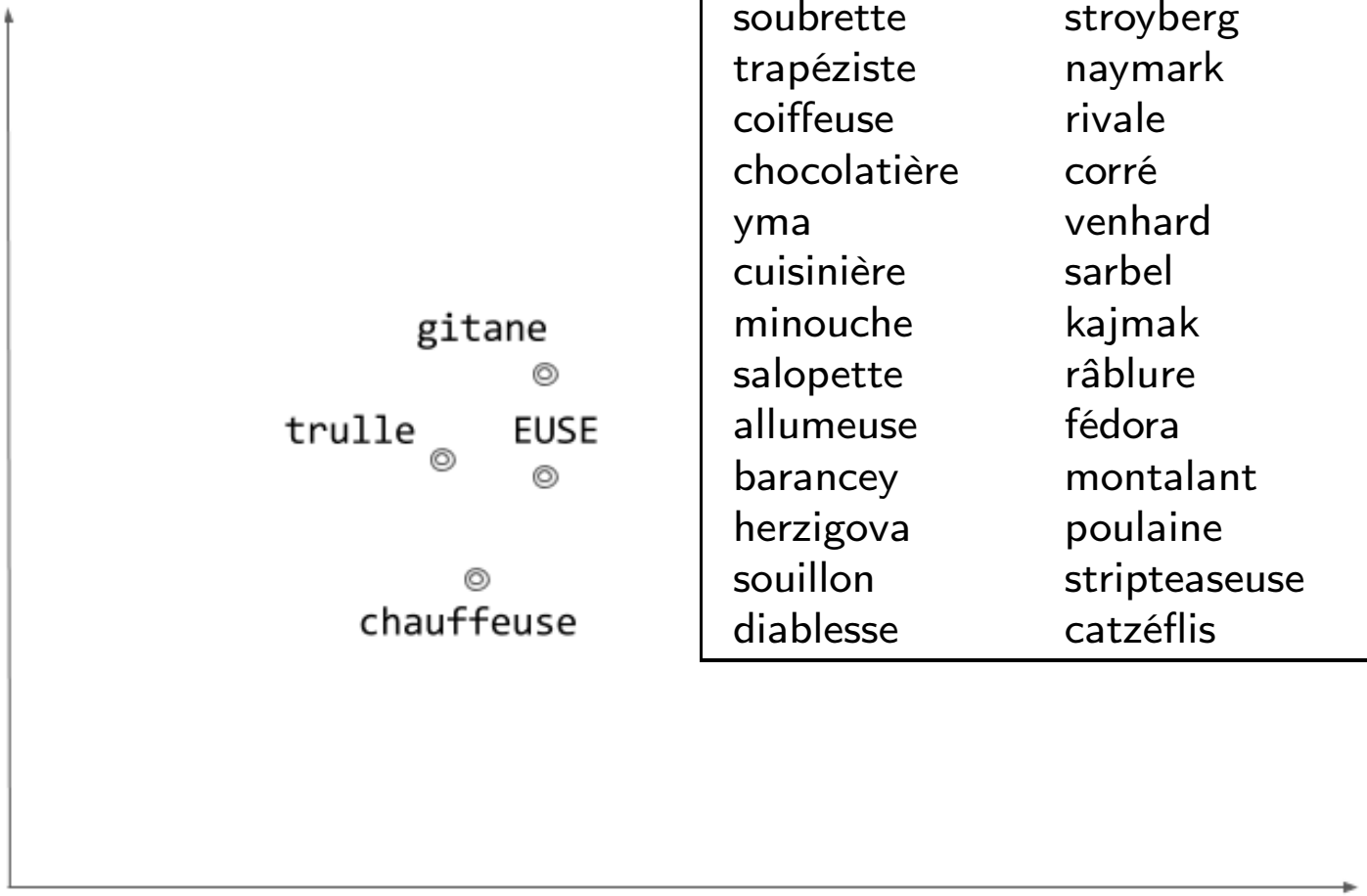
professeure	gogean	dartonne
co-fondatrice	desmarais-	warmus
cofondatrice	rondeau	présentatrice
herzigova	assistante	bourgeois-leclerc
directrice	mageina	tonietti
pharmacienne	musicienne	otternaud
traductrice	vyghen	directrice-
chercheure	ingénieure	adjointe
saint-lucienne	gérante	guirous
venhard	mammamia	saller
éducatrice	spaziani	sculptrice
co-directrice	anska	tshiteya
fondatrice	bouhenni	naymark
laury	slávka	écrivaine
conférencière	joano	rajskub
comédienne	séménoff	pomfresh
blogueuse	herzigová	fadeïeva
	shrier	



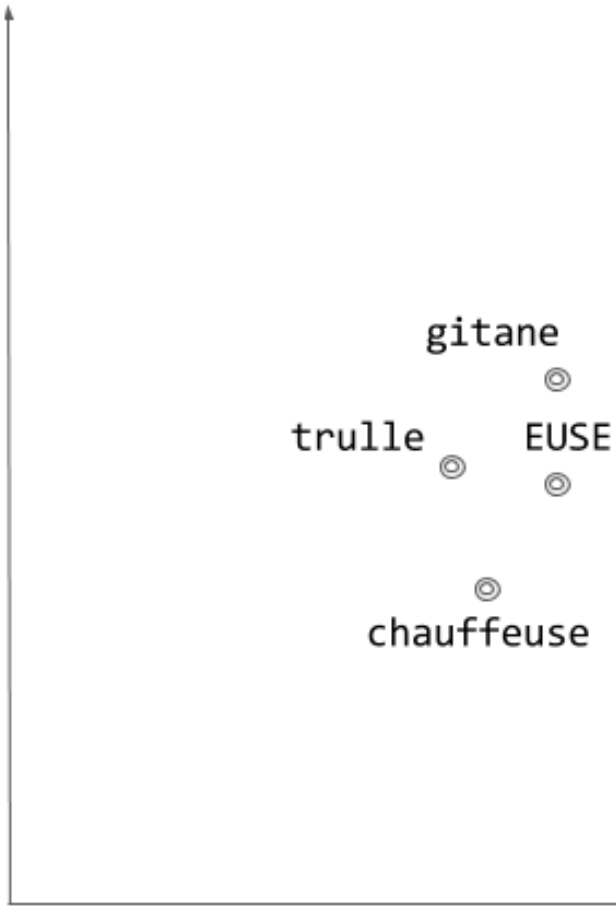
professeure	gogean	dartonne
co-fondatrice	desmarais-	warmus
cofondatrice	rondeau	présentatrice
herzigova	assistante	bourgeois-leclerc
directrice	mageina	tonietti
pharmacienne	musicienne	otternaud
traductrice	vyghen	directrice-
chercheure	ingénieure	adjointe
saint-lucienne	gérante	guirous
venhard	mammamia	saller
éducatrice	spaziani	sculptrice
co-directrice	anska	tshiteya
fondatrice	bouhenni	naymark
laury	slávka	écrivaine
conférencière	joano	rajskub
comédienne	séménoff	pomfresh
blogueuse	herzigová	fadeïeva
	shrier	



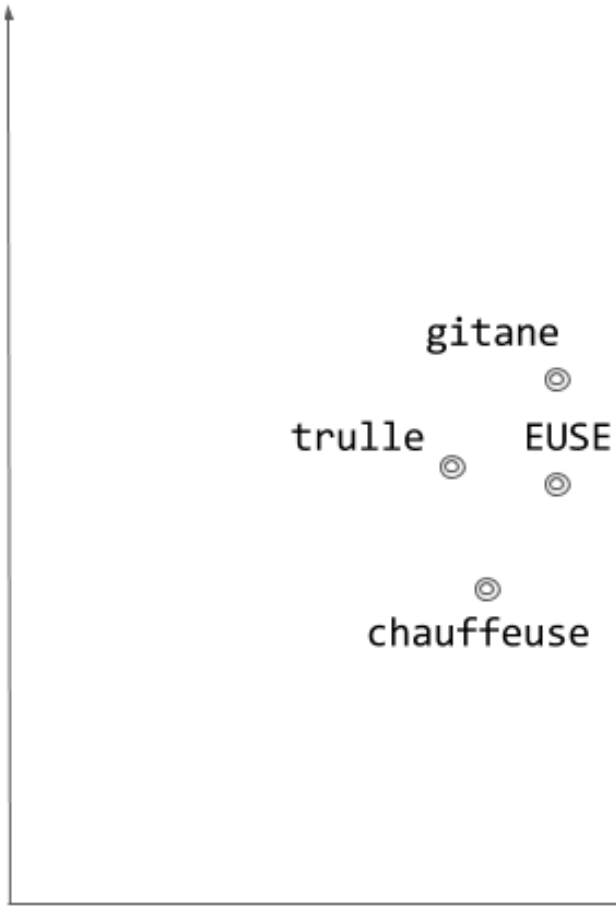
professeure	gogean	dartonne
co-fondatrice	desmarais-	warmus
cofondatrice	rondeau	**présentatrice**
herzigova	**assistante**	bourgeois-leclerc
directrice	mageina	tonietti
pharmacienne	**musicienne**	otternaud
traductrice	vyghen	**directrice-**
chercheure	**ingénieure**	**adjointe**
saint-lucienne	**gérante**	guirous
venhard	mammamia	saller
éducatrice	spaziani	**sculptrice**
co-directrice	anska	tshiteya
fondatrice	bouhenni	naymark
laury	slávka	**écrivaine**
conférencière	joano	rajskub
comédienne	séménoff	pomfresh
blogueuse	herzigová	fadeïeva
	shrier	



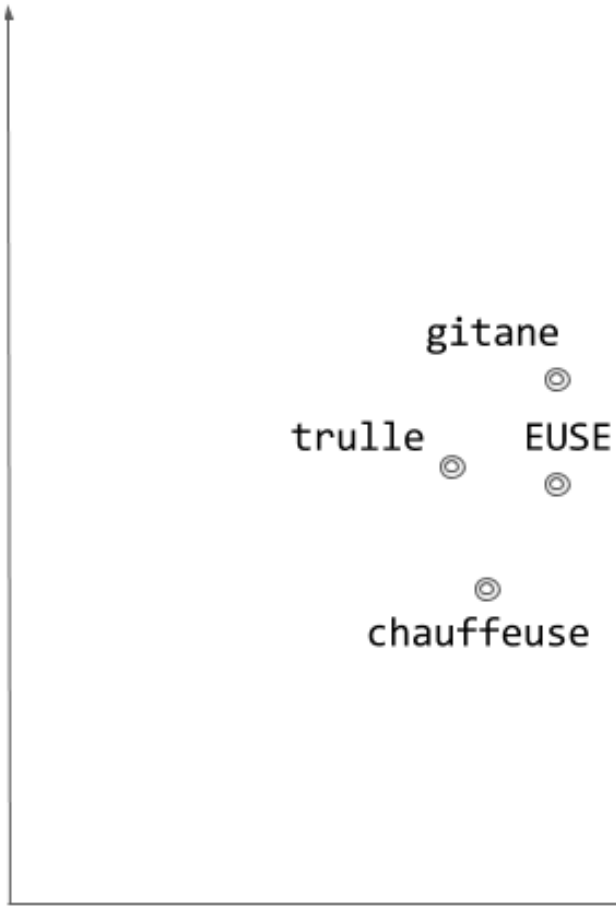
gitane	cochonne	mini-jupe
trulle	vericel	rosine
chauffeuse	serveuse	mariée
manucure	sorokina	ptereleotris
soubrette	stroyberg	tallier
trapéziste	naymark	irma
coiffeuse	rivale	suffel
chocolatière	corré	cover-girl
yma	venhard	épicière
cuisinière	sarbel	marie-olivier
minouche	kajmak	javotte
salopette	râblure	kerny
allumeuse	fédora	basquaise
barancey	montalant	emilienne
herzigova	poulaine	estragnat
souillon	stripteaseuse	tigresse
diablesse	catzéflis	



gitane	cochonne	mini-jupe
trulle	vericel	rosine
chauffeuse	serveuse	mariée
manucure	sorokina	ptereleotris
soubrette	stroyberg	tallier
trapéziste	naymark	irma
coiffeuse	rivale	suffel
chocolatière	corré	cover-girl
yma	venhard	épicière
cuisinière	sarbel	marie-olivier
minouche	kajmak	javotte
salopette	râblure	kerny
allumeuse	fédora	basquaise
barancey	montalant	emilienne
herzigova	poulaine	estragnat
souillon	stripteaseuse	tigresse
diablesse	catzéflis	



gitane	cochonne	mini-jupe
trulle	vericel	rosine
chauffeuse	**serveuse**	mariée
manucure	sorokina	ptereleotris
soubrette	stroyberg	tallier
trapéziste	naymark	irma
coiffeuse	rivale	suffel
chocolatière	corré	**cover-girl**
yma	venhard	**épicière**
cuisinière	sarbel	marie-olivier
minouche	kajmak	javotte
salopette	râblure	kerny
allumeuse	fédora	basquaise
barancey	montalant	emilienne
herzigova	poulaine	estragnat
souillon	stripteaseuse	tigresse
diablesse	catzéflis	



gitane
trulle
chauffeuse
manucure
soubrette
trapéziste
coiffeuse
chocolatière
yma
cuisinière
minouche
salopette
allumeuse
barancey
herzigova
souillon
diablesse

cochonne
vericel
serveuse
sorokina
stroyberg
naymark
rivale
corré
venhard
sarbel
kajmak
râblure
fédora
montalant
poulaine
stripteaseuse
catzéflis

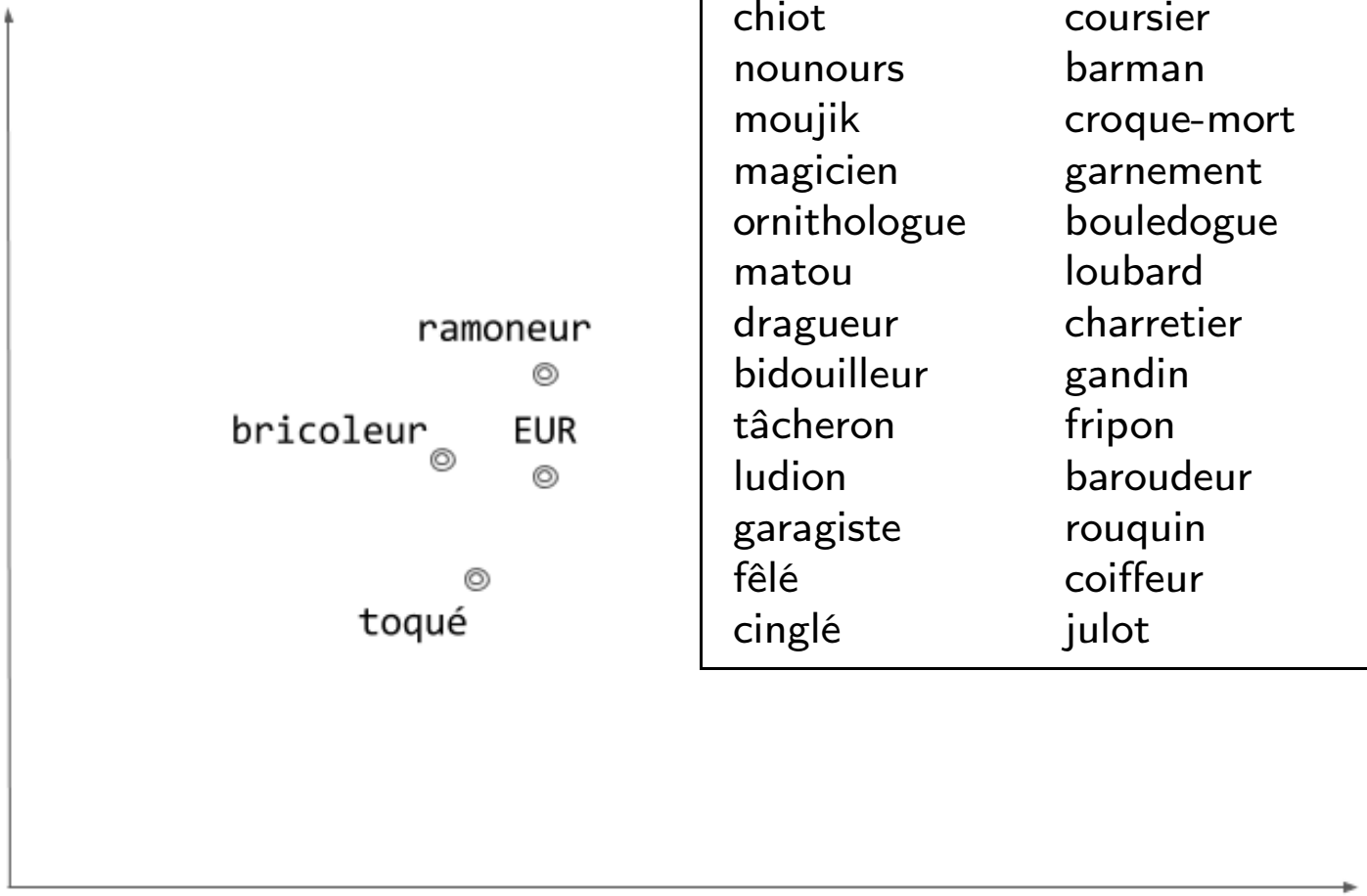
mini-jupe
rosine
mariée
ptereleotris
tallier
irma
suffel
cover-girl
épicière
marie-olivier
javotte
kerny
basquaise
emilienne
estragnat
tigresse

Discussion

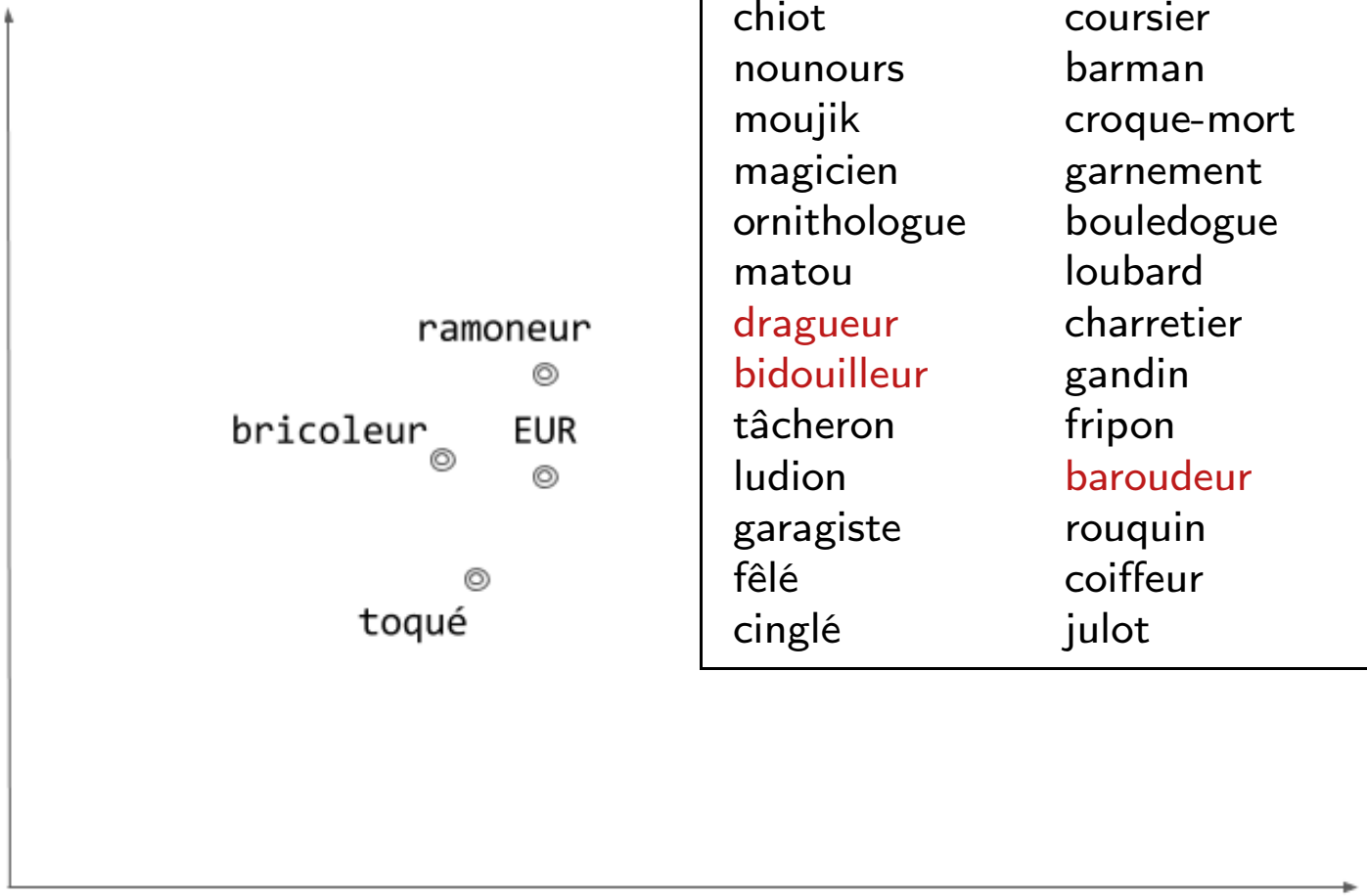
Dérivés prototypiques en *-eur*, *-euse* et *-rice* bien distincts (Caliskan et al., 2017; Kisselew et al., 2015)

- ***-eur*** : agents masculins et instruments
- ***-rice*** : agents féminins et humains de sexe féminin
- ***-euse*** : agents féminins, humains de sexe féminin, concepts et connotations associés au sexe féminin

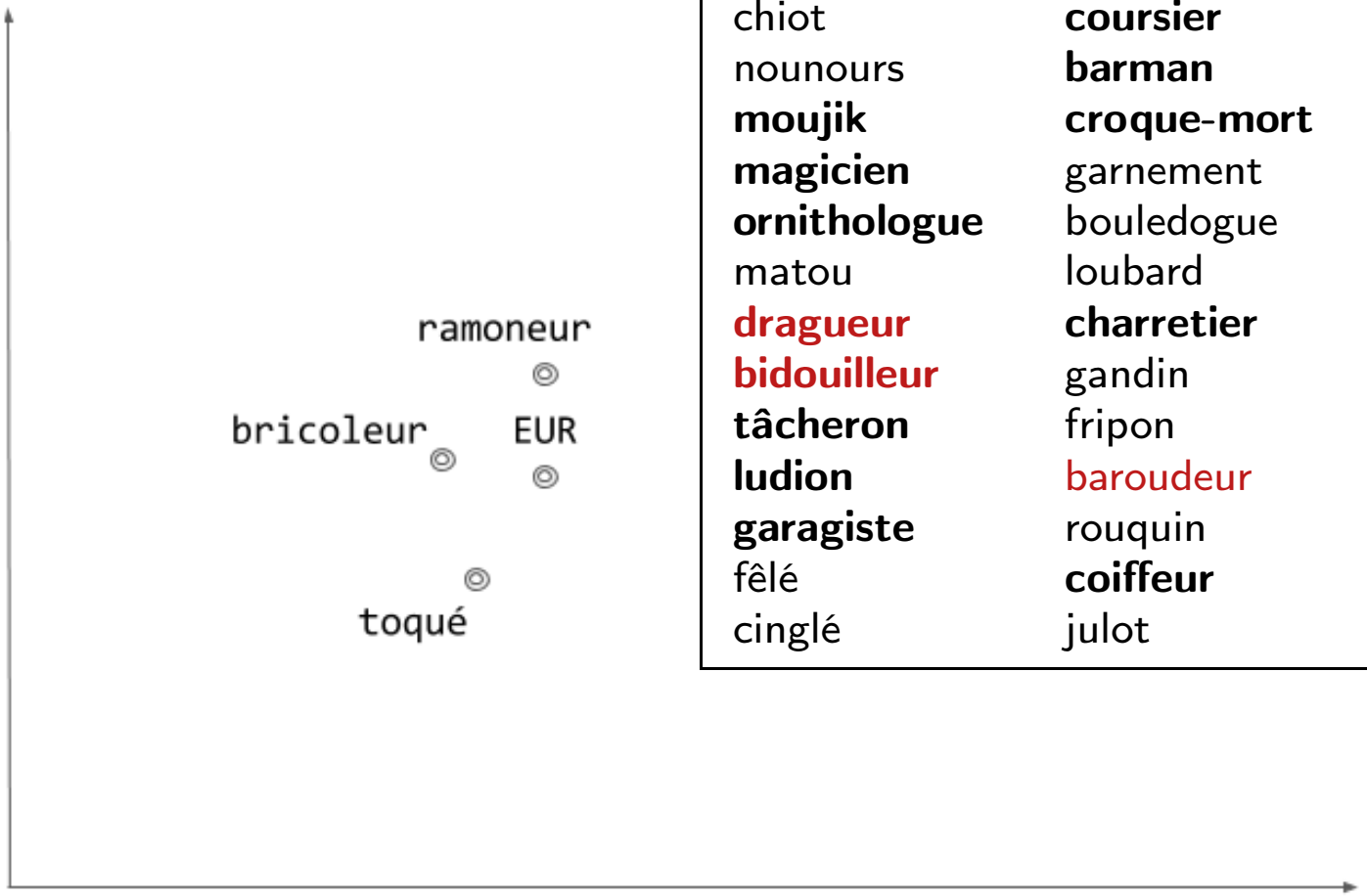
Liés à la caractérisation des femmes dans Wikipedia (Wagner et al., 2015) mais pas limités à *Wikipedia*



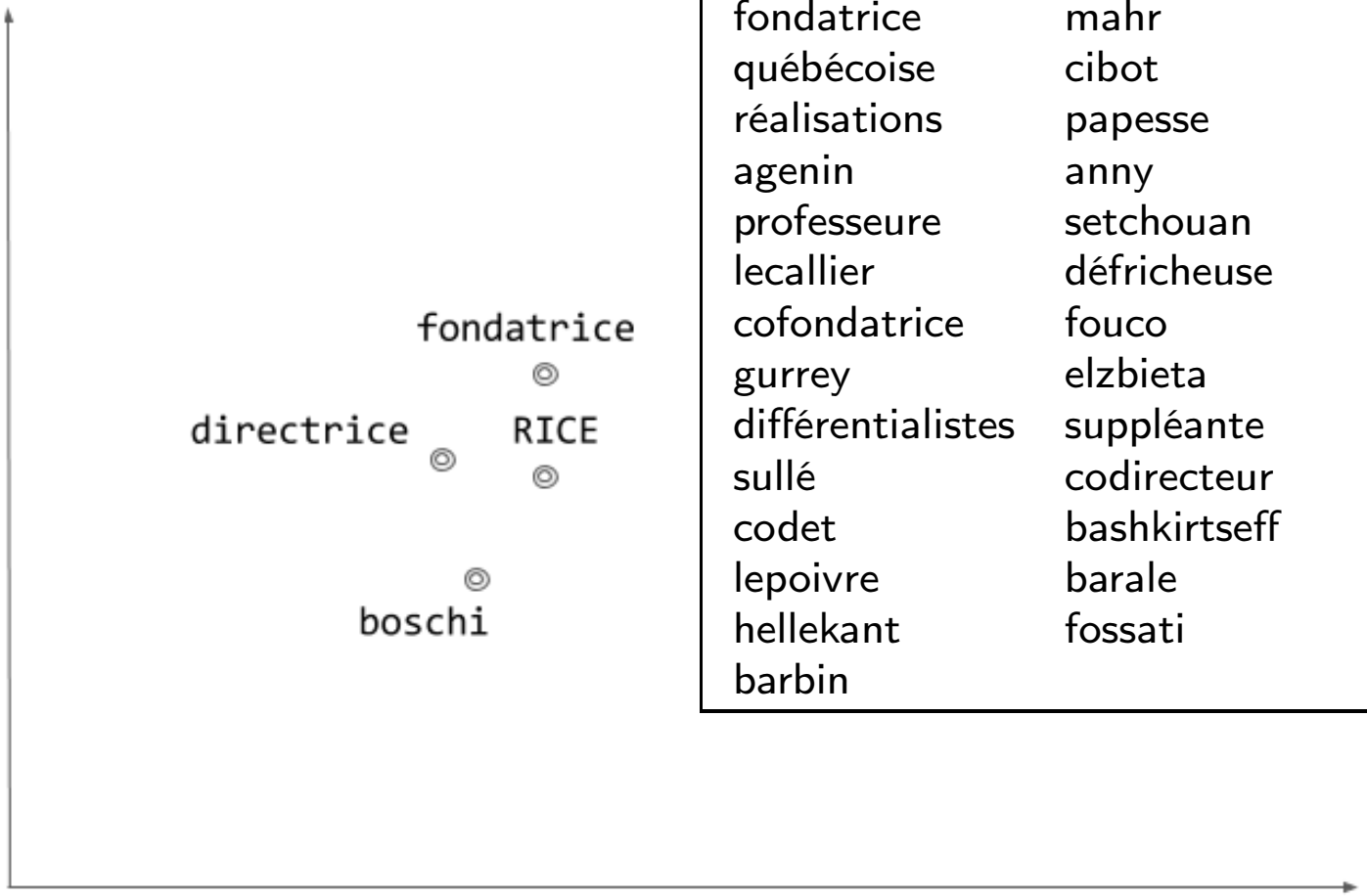
ramoneur	comparse	boxeur
bricoleur	imitateur	arnaqueur
toqué	frelon	malfrat
alchimiste	aventurier	voyou
chiot	coursier	écuyer
nounours	barman	prestidigitateur
moujik	croque-mort	moussaillon
magicien	garnement	cuisinier
ornithologue	bouledogue	sarret
matou	loubard	puncheur
dragueur	charretier	fêtard
bidouilleur	gandin	camelot
tâcheron	fripon	affabulateur
ludion	baroudeur	braconnier
garagiste	rouquin	canari
fêlé	coiffeur	garçon
cinglé	julot	



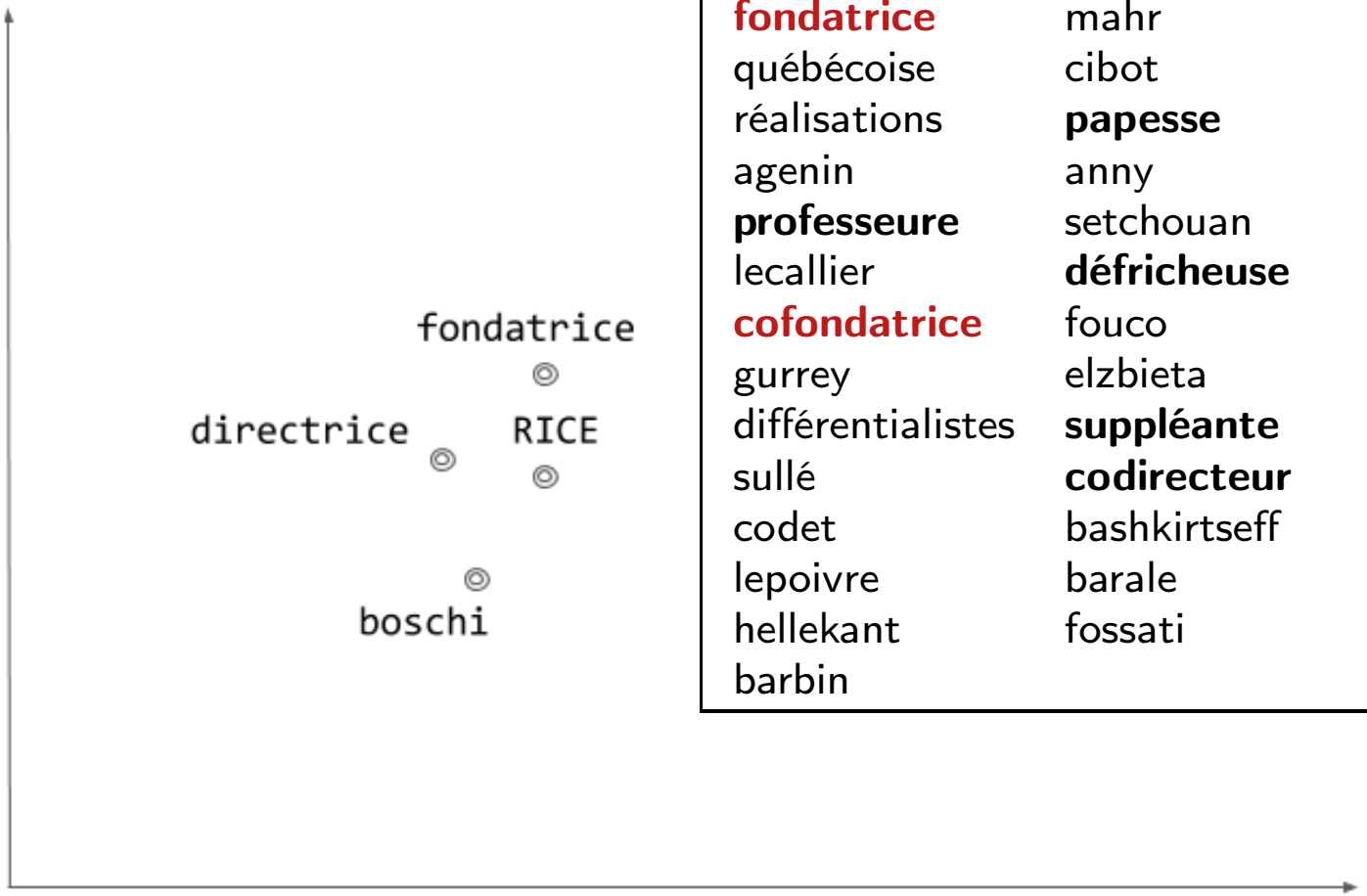
ramoneur	comparse	boxeur
bricoleur	imitateur	arnaqueur
toqué	frelon	malfrat
alchimiste	aventurier	voyou
chiot	coursier	écuyer
nounours	barman	prestidigitateur
moujik	croque-mort	moussaillon
magicien	garnement	cuisinier
ornithologue	bouledogue	sarret
matou	loubard	puncheur
dragueur	charretier	fêtard
bidouilleur	gandin	camelot
tâcheron	fripon	affabulateur
ludion	baroudeur	braconnier
garagiste	rouquin	canari
fêlé	coiffeur	garçon
cinglé	julot	



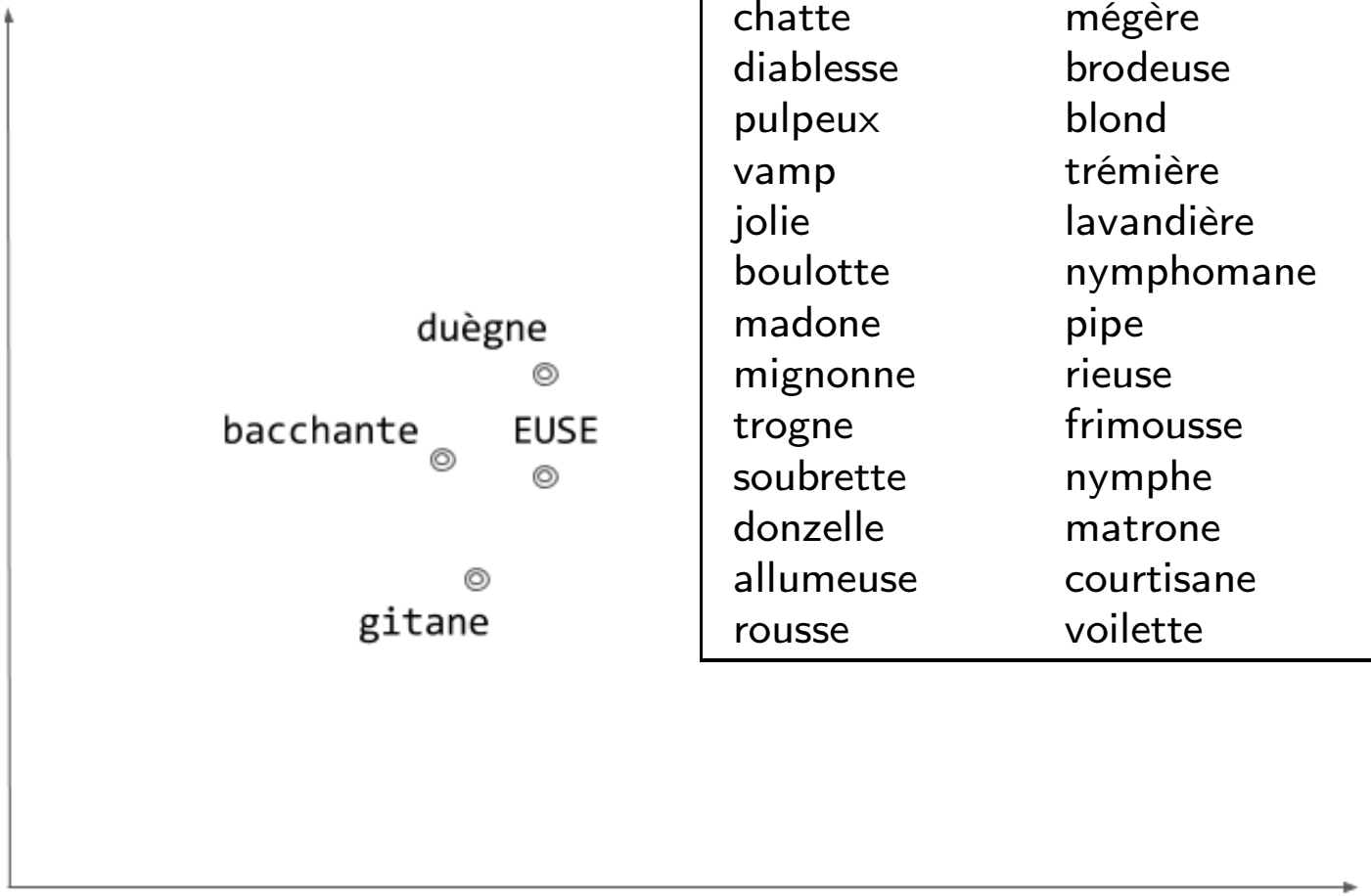
ramoneur	comparse	**boxeur**
bricoleur	**imitateur**	**arnaqueur**
toqué	frelon	**malfrat**
alchimiste	**aventurier**	voyou
chiot	**coursier**	**écuyer**
nounours	**barman**	**prestidigitateur**
moujik	**croque-mort**	**moussaillon**
magicien	garnement	**cuisinier**
ornithologue	bouledogue	sarret
matou	loubard	**puncheur**
dragueur	**charretier**	fêtard
bidouilleur	gandin	camelot
tâcheron	fripon	affabulateur
ludion	baroudeur	**braconnier**
garagiste	rouquin	canari
fêlé	**coiffeur**	garçon
cinglé	julot	



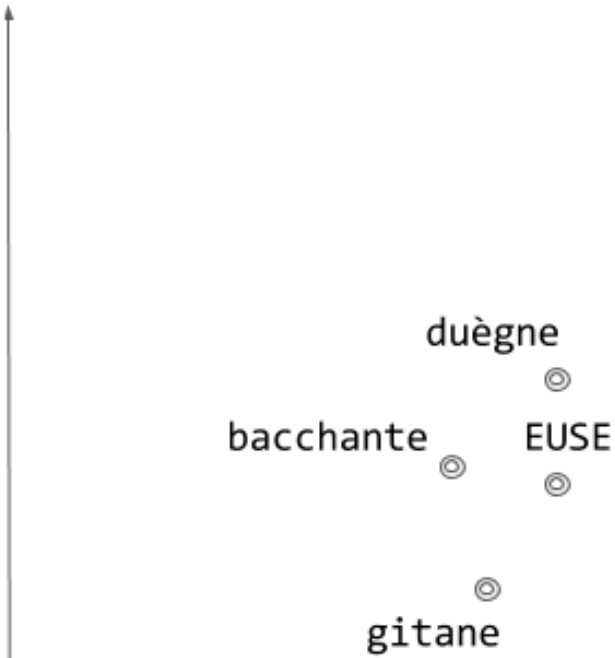
fondatrice	ambrosi	salvaire
directrice	solange	gourfink
boschi	bécue	diaconesse
présidente-	experte	lucienne
fondatrice	mahr	mémoriale
québécoise	cibot	lisette
réalisations	papesse	pumain
agenin	anny	woringer
professeure	setchouan	wouters
lecallier	défricheuse	busq
cofondatrice	fouco	swenson
gurrey	elzbieta	ballis
différentialistes	suppléante	aubriot
sullé	codirecteur	beaurain
codet	bashkirtseff	tinguy
lepoivre	barale	lausseur
hellekant	fossati	
barbin		



fondatrice	ambrosi	salvaire
directrice	solange	gourfink
boschi	bécue	**diaconesse**
présidente-	**experte**	lucienne
fondatrice	mahr	mémoriale
québécoise	cibot	lisette
réalisations	**papesse**	pumain
agenin	anny	woringer
professeure	setchouan	wouters
lecallier	**défricheuse**	busq
cofondatrice	fouco	swenson
gurrey	elzbieta	ballis
différentialistes	**suppléante**	aubriot
sullé	**codirecteur**	beaurain
codet	bashkirtseff	tinguy
lepoivre	barale	lausseur
hellekant	fossati	
barbin		



duègne	silhouetter	ballerine
bacchante	blonde	servante
gitane	garce	femme-oiseau
ravissant	adorable	midinette
chatte	mégère	naïade
diablesse	brodeuse	espiègle
pulpeux	blond	blondeur
vamp	trémière	almée
jolie	lavandière	guenon
boulotte	nymphomane	pomponner
madone	pipe	femme-enfant
mignonne	rieuse	teigne
trogne	frimousse	minauder
soubrette	nymphe	strip-teaseuse
donzelle	matrone	pépée
allumeuse	courtisane	citrouille
rousse	voilette	



duègne	silhouetter	ballerine
bacchante	blonde	servante
gitane	garce	femme-oiseau
ravissant	adorable	midinette
chatte	mégère	naïade
diablesse	brodeuse	espiègle
pulpeux	blond	blondeur
vamp	trémière	almée
jolie	**lavandière**	guenon
boulotte	nymphomane	pomponner
madone	pipe	femme-enfant
mignonne	rieuse	teigne
trogne	frimousse	minauder
soubrette	nymphe	**strip-teaseuse**
donzelle	matrone	pépée
allumeuse	**courtisane**	citrouille
rousse	voilette	

Plan

Pourquoi la nominalisation en *-age*, *-ion* et *-ment*

- Nominalisation
  - Création d'un nom d'action déverbal
    - *ramoner* - *ramonage*
  - Suffixes *-age*, *-ion* et *-ment* les plus utilisés (Fradin, 2014)
  - Une même glose : *Action de V*
    - *ramonage* : action de ramoner
  - Plusieurs interprétations (Huyghe, 2014; Balvet et al., 2011)
    - Activité : *jardinage* - *construction* - *creusement*
    - Événement : *accrochage* - *construction* - *débarquement*
    - Objet : *barrage* - *construction* - *amendement*
    - État : *rééquilibrage* - *déception* - *mécontentement*

- Concurrence suffixale
  - *déambulage* (0) - *déambulation* (91) - *déambulement* (0)
  - *pavage* (487) - *pavement* (449)
    - *quand on réalise une projection de la rose de la façade sur le* ***pavement***
    - *Des repères fixés de point en point sur le* ***pavage*** *permettent un guidage des touristes*

- Qu'est-ce qui distingue ces suffixes ?

Propriétés sémantiques (1)

- Nature sémantique des arguments du verbe (Martin, 2010; Fradin, 2014; Dubois, 1962)
  - *décollage* vs *décollement*
- Télicité du verbe (Martin, 2010)
  - *construire* vs *jardiner*
- Longueur de la chaîne événementielle dénotée par le nom (Martin, 2010)
  - *gonflage* vs *gonflement*
- Incrémentalité de l'action dénotée par le nom (Martin, 2010)
  - *plissage* vs *plissement*
- Domaine ontologie (Dubois, 1962; Martin, 2010)
  - *-age* pour le domaine physique et pour les opérations industrielles (*abattage*)
  - *-ion* au lexique scientifique et technique (Dubois, 1962)
  - *-ment* pour le domaine abstrait (*abattement*)

Propriétés formelles et syntaxiques

- Transitivité du verbe (Dubois, 1962; Fradin, 2014)
  - *vérifier* vs *voyager*
- Base (Fradin, 2014)
  - Base savante pour *-ion* (*production*)
  - Base populaire pour *-age* et *-ment* (*abattage*)

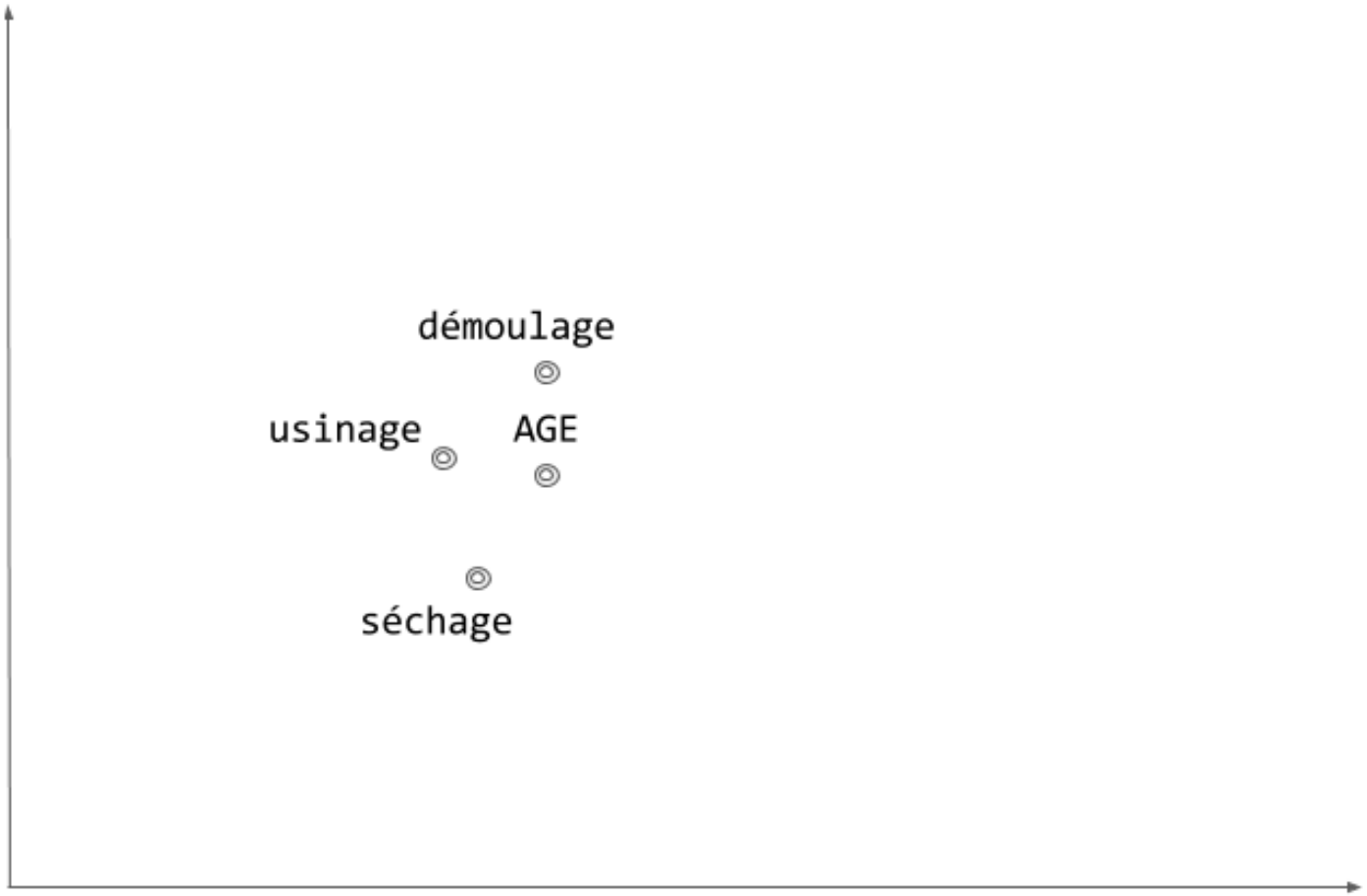
Vers une nominalisation prototypique

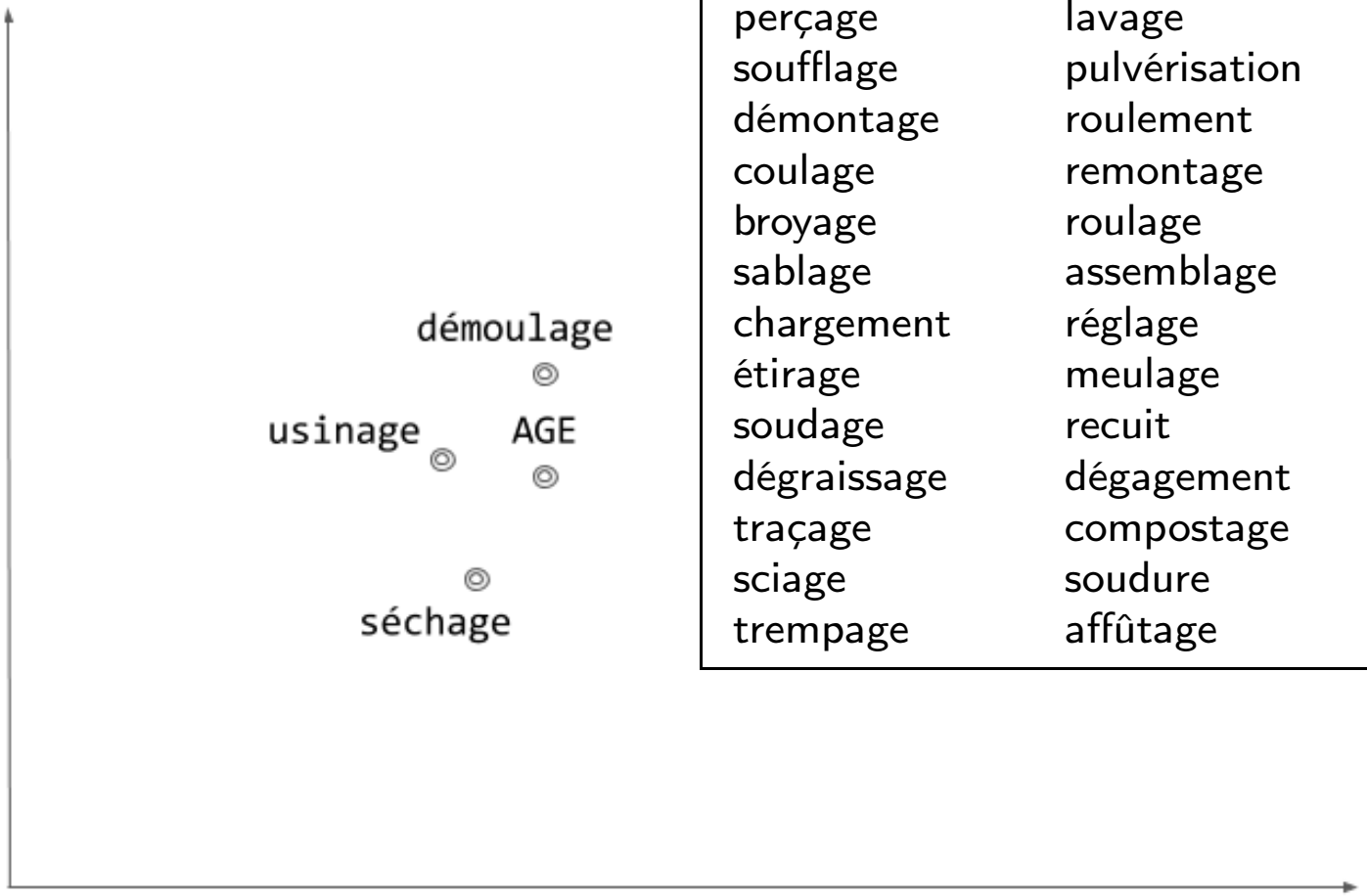
- Reprise de l'expérience précédente mais pour les noms d'action en *-age*, *-ion* et *-ment*

$$\overrightarrow{SUFF} = \frac{\overrightarrow{\text{Nsuff}_1} + \overrightarrow{\text{Nsuff}_2} + ... + \overrightarrow{\text{Nsuff}_n}}{n}$$

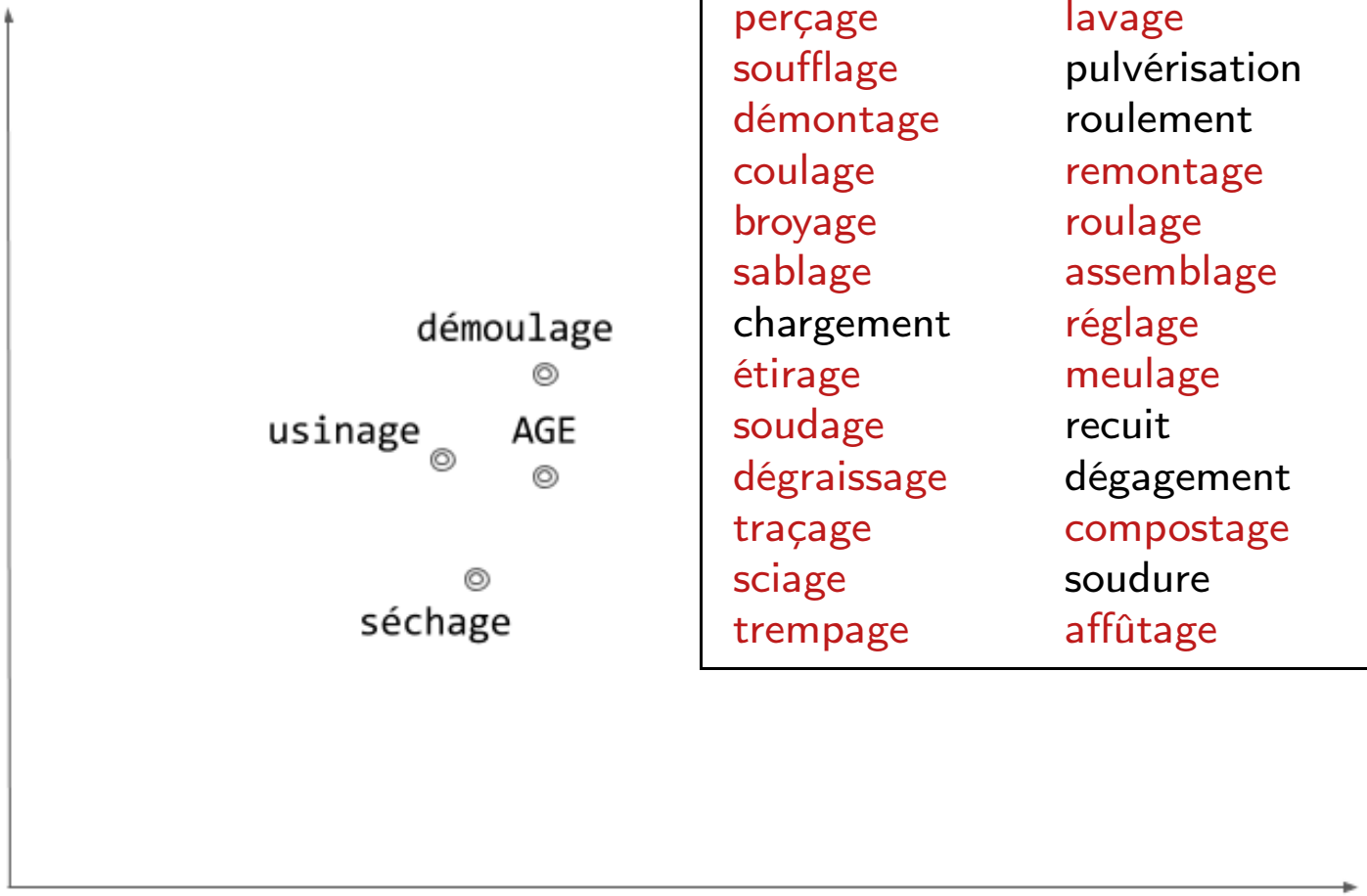
- Paramètres
  - Corpus : *Wikipedia* vs *LM10*
  - Nombre de dimensions : *100* vs *300*
  - Algorithme : *CBOW* vs *Skip-gram*

	-age	*-ion*	*-ment*
Wikipedia	707	1635	592
LM10	563	1507	561

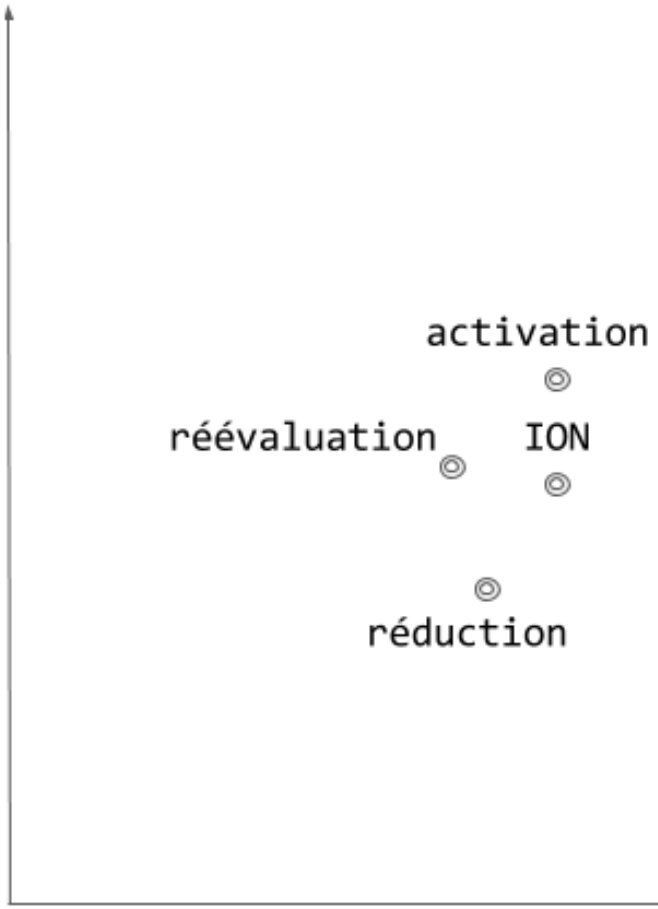




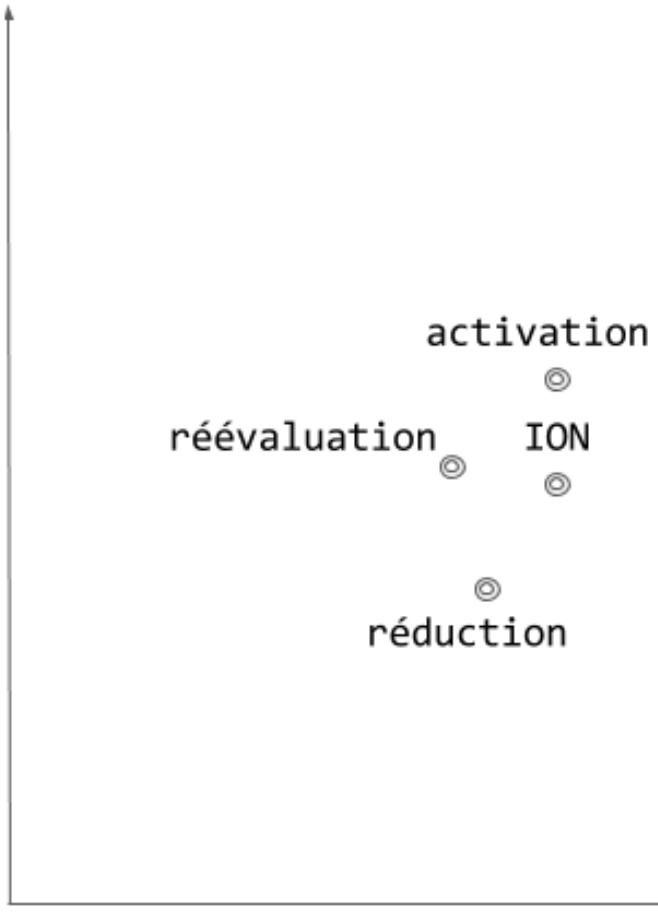
démoulage	polissage	salage
usinage	vissage	désinfection
séchage	nettoyage	foulage
remplissage	gonflage	cuivrage
perçage	lavage	enrobage
soufflage	pulvérisation	vidange
démontage	roulement	refroidissement
coulage	remontage	étanchéité
broyage	roulage	clouage
sablage	assemblage	décapage
chargement	réglage	rechargement
étirage	meulage	stockage
soudage	recuit	grattage
dégraissage	dégagement	rinçage
traçage	compostage	dégivrage
sciage	soudure	brasage
trempage	affûtage	



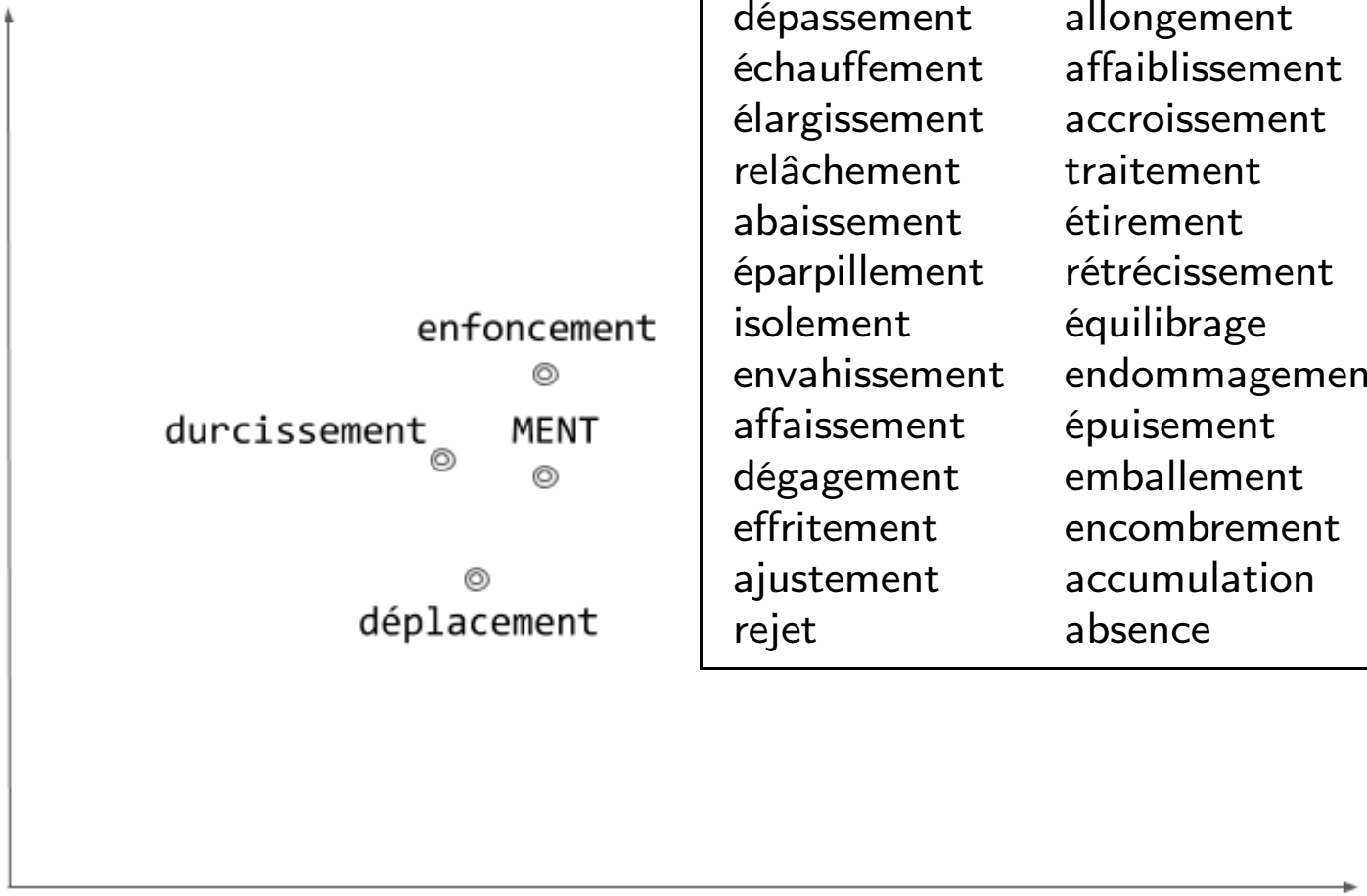
démoulage	polissage	salage
usinage	vissage	désinfection
séchage	nettoyage	foulage
remplissage	gonflage	cuivrage
perçage	lavage	enrobage
soufflage	pulvérisation	vidange
démontage	roulement	refroidissement
coulage	remontage	étanchéité
broyage	roulage	clouage
sablage	assemblage	décapage
chargement	réglage	rechargement
étirage	meulage	stockage
soudage	recuit	grattage
dégraissage	dégagement	rinçage
traçage	compostage	dégivrage
sciage	soudure	brasage
trempage	affûtage	



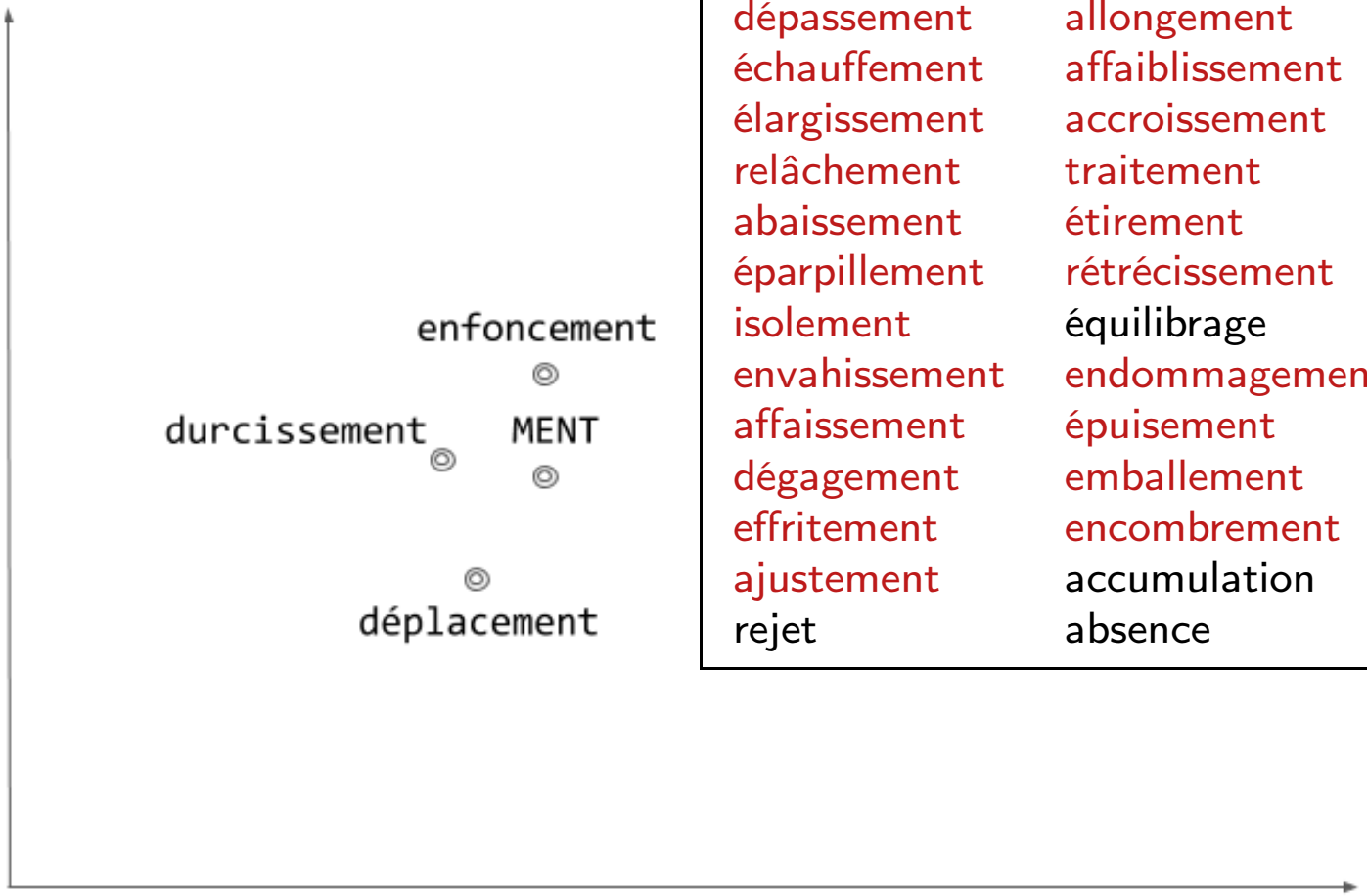
activation	stimulation	intervention
réévaluation	dilatation	limitation
réduction	généralisation	appropriation
simplification	différenciation	régénération
dégradation	altération	formulation
détérioration	action	imputation
assimilation	mutation	inhibition
acceptation	dispersion	prolifération
utilisation	complexification	perception
modification	dénaturation	définition
transformation	homogénéisation	analyse
manipulation	vérification	constatation
application	réaction	dissociation
évaluation	survenue	actualisation
dilution	coupure	accumulation
détermination	compréhension	mesure
surcharge	fixation	



activation	stimulation	intervention
réévaluation	dilatation	limitation
réduction	généralisation	appropriation
simplification	différenciation	régénération
dégradation	altération	formulation
détérioration	action	imputation
assimilation	mutation	inhibition
acceptation	dispersion	prolifération
utilisation	complexification	perception
modification	dénaturation	définition
transformation	homogénéisation	analyse
manipulation	vérification	constatation
application	réaction	dissociation
évaluation	survenue	actualisation
dilution	coupure	accumulation
détermination	compréhension	mesure
surcharge	fixation	



enfoncement	écoulement	remplissage
durcissement	dysfonctionnement	prélèvement
déplacement	ralentissement	inconfort
blocage	basculement	tassement
dépassement	allongement	ensablement
échauffement	affaiblissement	éloignement
élargissement	accroissement	renforcement
relâchement	traitement	utilisation
abaissement	étirement	étalement
éparpillement	rétrécissement	engorgement
isolement	équilibrage	usure
envahissement	endommagement	redémarrage
affaissement	épuisement	lessivage
dégagement	emballement	décollement
effritement	encombrement	gonflement
ajustement	accumulation	colmatage
rejet	absence	



enfoncement	écoulement	remplissage
durcissement	dysfonctionnement	prélèvement
déplacement	ralentissement	inconfort
blocage	basculement	tassement
dépassement	allongement	ensablement
échauffement	affaiblissement	éloignement
élargissement	accroissement	renforcement
relâchement	traitement	utilisation
abaissement	étirement	étalement
éparpillement	rétrécissement	engorgement
isolement	équilibrage	usure
envahissement	endommagement	redémarrage
affaissement	épuisement	lessivage
dégagement	emballement	décollement
effritement	encombrement	gonflement
ajustement	accumulation	colmatage
rejet	absence	

Pourcentage de voisins porteurs du suffixe

		-age	*-ion*	*-ment*
Wikipedia	100	76%	88%	74%
	300	68%	82%	78%
LM10	100	56%	84%	82%
	300	54%	92%	84%

		-age	*-ion*	*-ment*
Wikipedia	CBOW	76%	88%	74%
	Skip-gram	44%	22%	12%
LM10	CBOW	56%	84%	82%
	Skip-gram	26%	32%	18%

Recouvrement des voisins

-age	*-ion*	*-ment*
36%	36%	17%

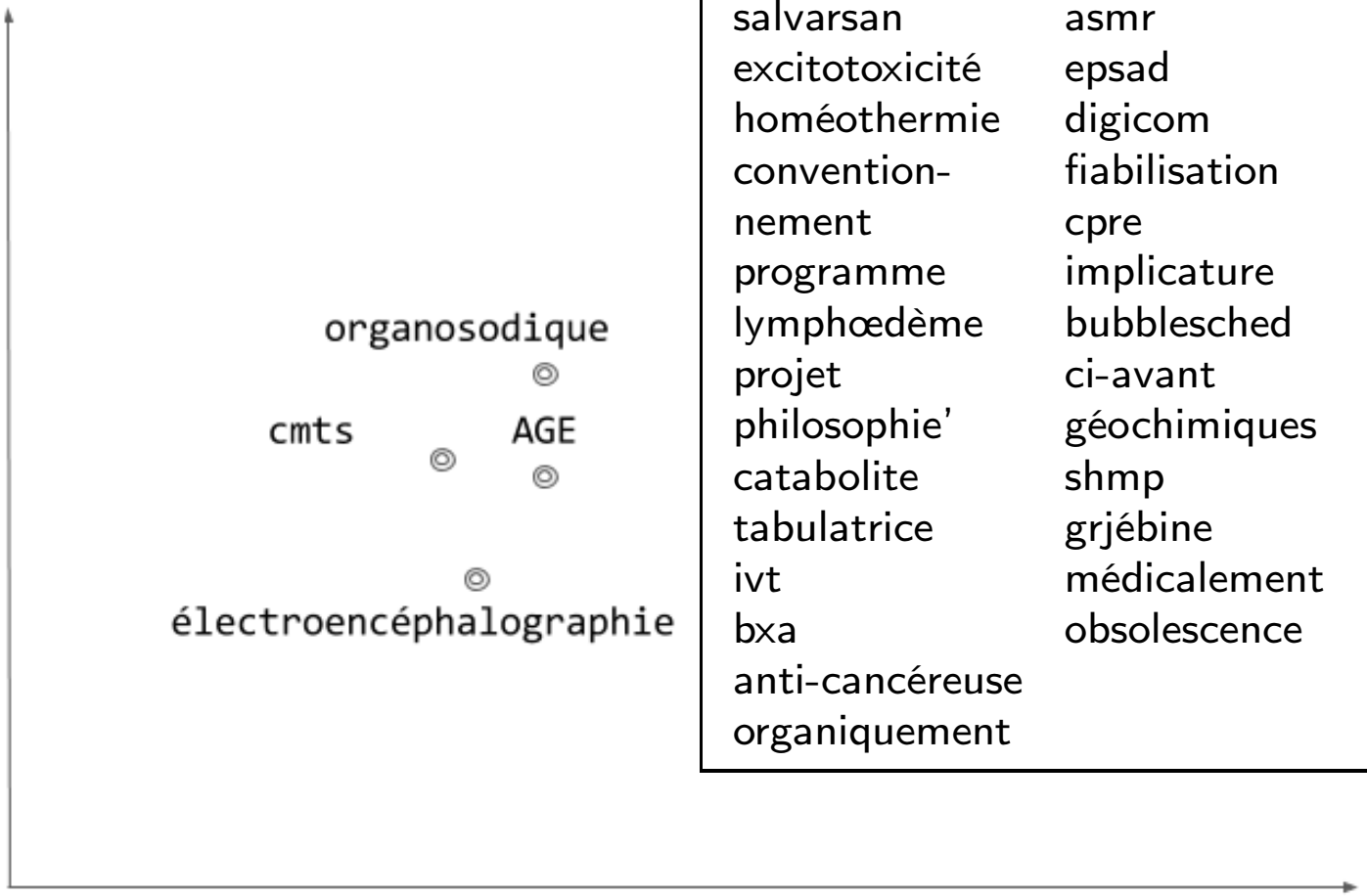
Table : Entre les corpus (CBOW 100 dimensions)

	-age	*-ion*	*-ment*
Wikipedia	82%	70%	86%
LM10	74%	68%	76%

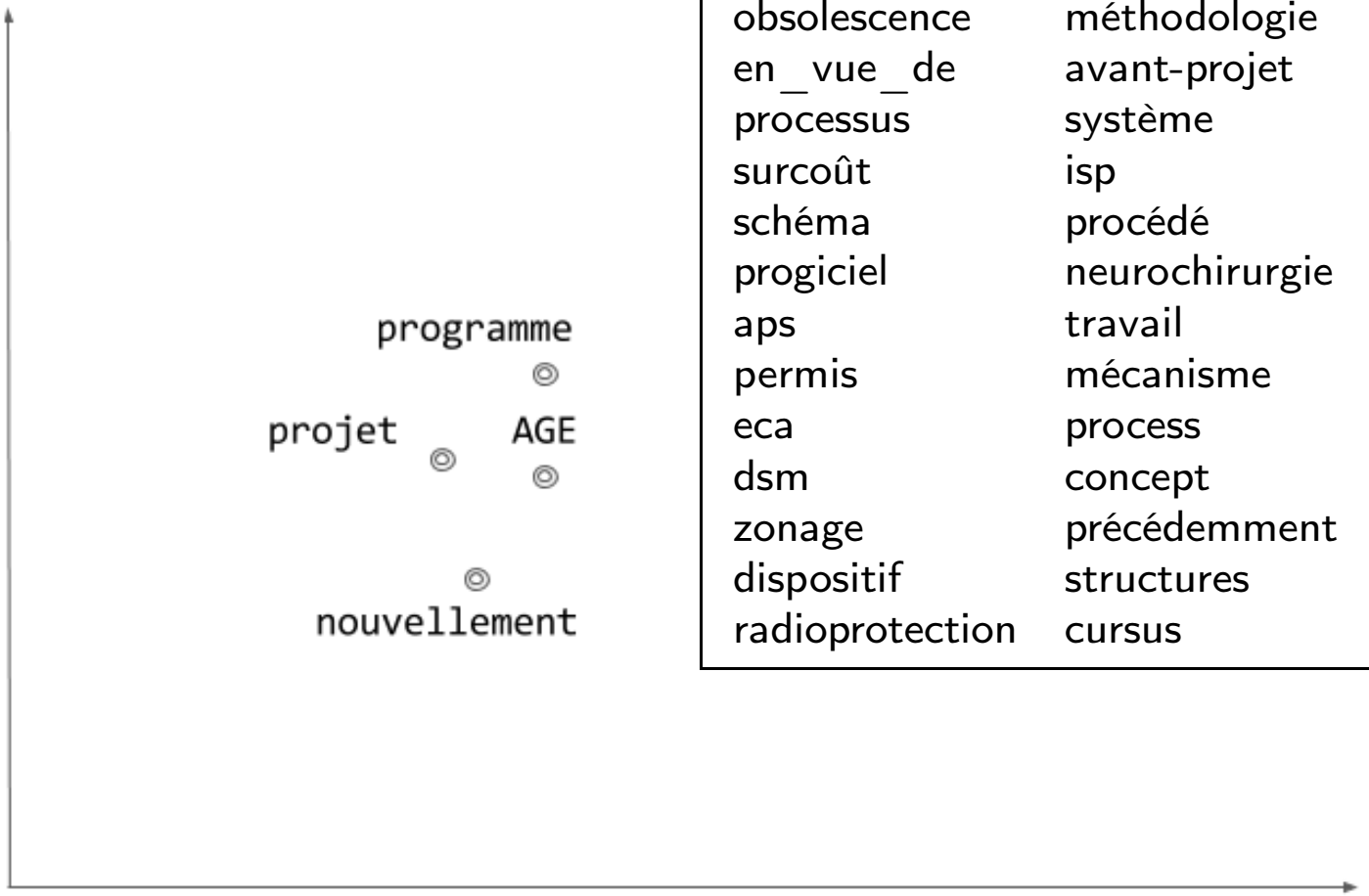
Table : Pour la variation du nombre de dimensions (CBOW)

	-age	*-ion*	*-ment*
Wikipedia	0%	0%	0%
LM10	0%	0%	0%

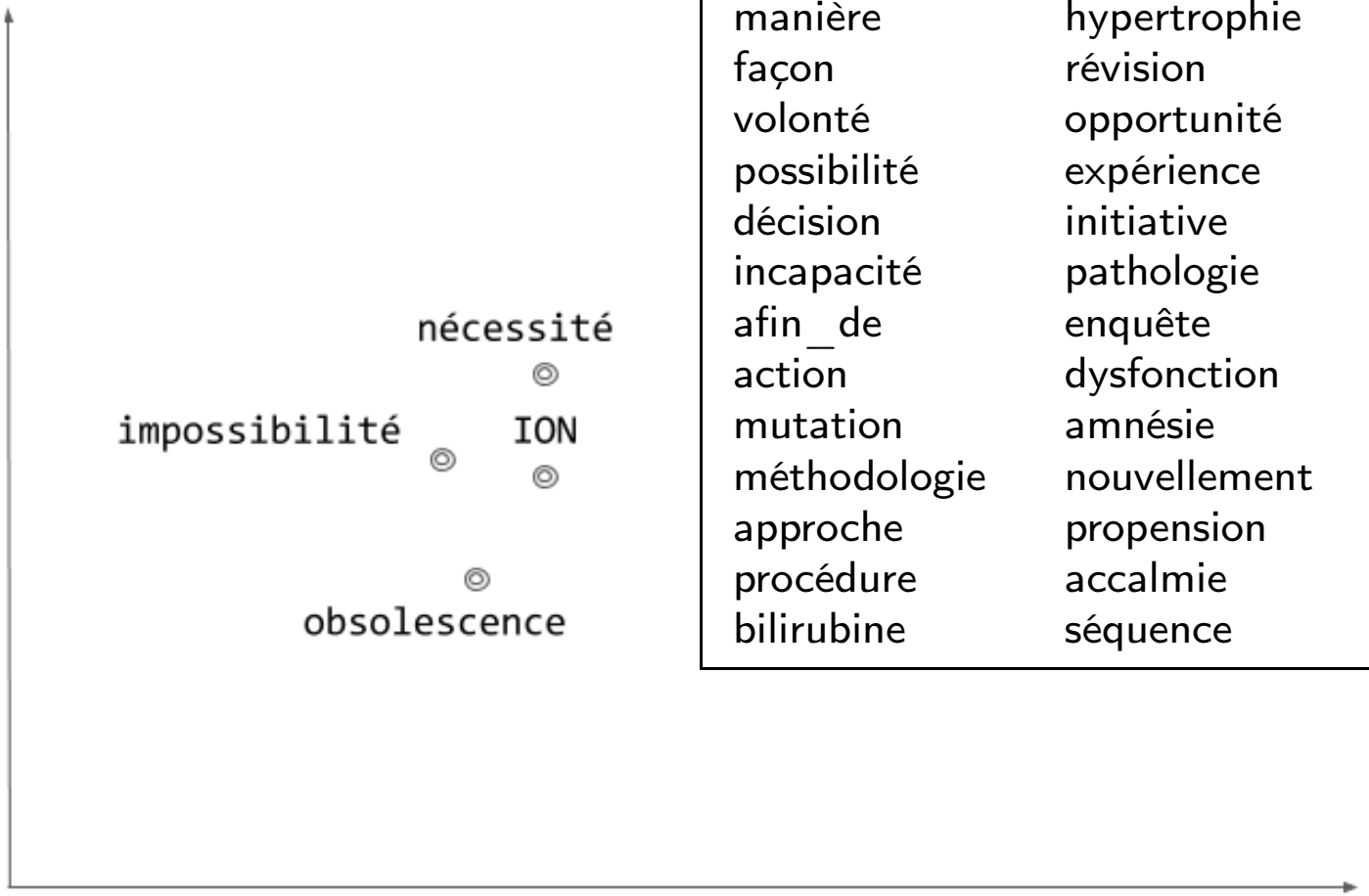
Table : Entre les deux algorithmes (100 dimensions)



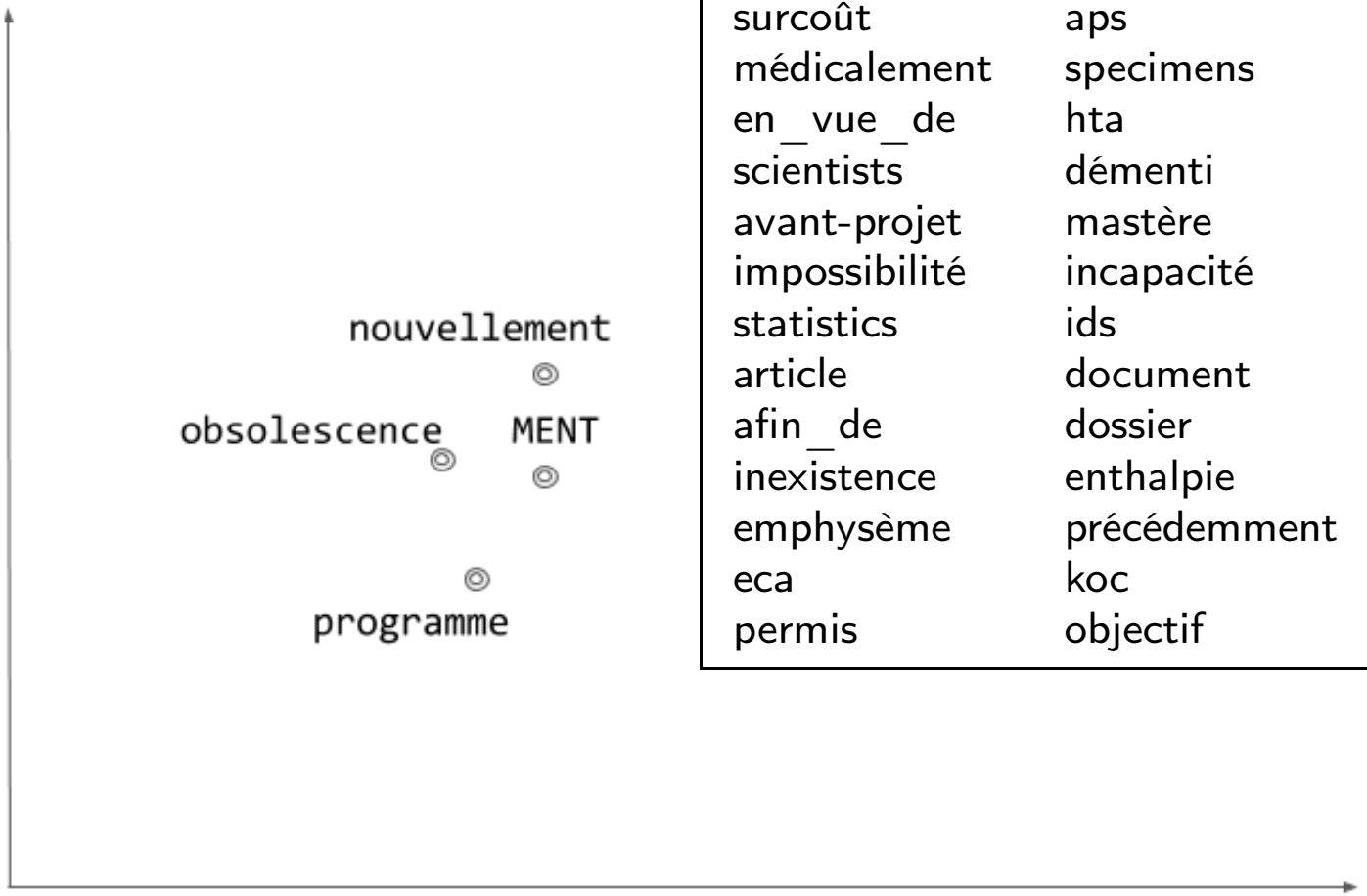
organosodique	vdi	ponts-rails
cmts	nouvellement	cng
électroencé-	microsystèmes	en_vue_de
phalographie	linstaller	r-d
salvarsan	asmr	proto-
excitotoxicité	epsad	industrialisation
homéothermie	digicom	entretient
convention-	fiabilisation	pamidronate
nement	cpre	hvs
programme	implicature	ppur
lymphœdème	bubblesched	lmcu
projet	ci-avant	processus
philosophie'	géochimiques	parfumeur-
catabolite	shmp	créateur
tabulatrice	grjébine	distributions
ivt	médicalement	exobiologie
bxa	obsolescence	gtk-doc
anti-cancéreuse		morphogènes
organiquement		



programme	statistics	a_posteriori
projet	afin_de	typage
nouvellement	ids	expertise
médicalement	mastère	codon
obsolescence	méthodologie	réseau
en_vue_de	avant-projet	correctif
processus	système	faisabilité
surcoût	isp	isae
schéma	procédé	hta
progiciel	neurochirurgie	logiciel
aps	travail	informatisation
permis	mécanisme	suivi
eca	process	tests
dsm	concept	métrologie
zonage	précédemment	développement
dispositif	structures	fallot
radioprotection	cursus	



nécessité	dépolarisation	stratégie
impossibilité	médicalement	orientation
obsolescence	cardiopathie	décroissance
démarche	solution	déviation
manière	hypertrophie	refonte
façon	révision	autorisation
volonté	opportunité	thèse
possibilité	expérience	structure
décision	initiative	dégénérescence
incapacité	pathologie	réévaluation
afin_de	enquête	complétion
action	dysfonction	transplantation
mutation	amnésie	proposition
méthodologie	nouvellement	analyse
approche	propension	cirrhose
procédure	accalmie	correction
bilirubine	séquence	



nouvellement	schéma	ap-hp
obsolescence	processus	dûment
programme	zonage	codon
projet	dsm	typage
surcoût	aps	évènement
médicalement	specimens	eeg
en_vue_de	hta	dépeuplement
scientists	démenti	ined
avant-projet	mastère	curriculum
impossibilité	incapacité	initiative
statistics	ids	dispositif
article	document	autisme
afin_de	dossier	renommage
inexistence	enthalpie	mémorandum
emphysème	précédemment	action
eca	koc	isp
permis	objectif	

Voisins partagés

- Par les trois suffixes : 4
  - *obsolescence*, *médicalement*, *nouvellement* et *afin_de*
- Par *-age* et *-ment* : 22
  - *programme*, *projet*, *dispositif*, *processus*, *schéma*...
- Par *-age* et *-ion* : 1
  - *méthodologie*
- Par *-ion* et *-ment* : 4
  - *impossibilité*, *incapacité*, *action* et *initiative*

Conclusions de l'expérience

Linguistiques

- Des régularités
- Distinction sémantique en plus de formelle
  - *-age* plus technique
  - *-ion* davantage sous-spécifié
  - *-ment* intermédiaire entre les deux
    - *-ion* comme opérateur
    - Dérivé en *-ion* comme opérateur ?

Méthodologiques

- Barycentres additif vs soustractif

Plan

Pour résumer

Premiers résultats encourageants

- Application de la SD à la morphologie
- Dérivés agentifs
  - ***-eur*** **:** agent et instrument
  - ***-rice*** **:** agent et féminin
  - ***-euse*** **:** féminin et péjoratif
- Nominalisation
  - ***-age*** **:** termes techniques et/ou spécifiques
  - ***-ion*** **:** termes sous-spécifiés
  - ***-ment*** **:** intermédiaire entre *-age* et *-ion*

Perspectives (1)

Pousser l'analyse de -*age*, -*ion* et -*ment*

- Compléter les expériences
  - Comparaison CBOW/Skip-gram
  - Extension aux modèles syntaxiques
- Mettre en regard des critères évoqués
  - L'étude de l'automatisation de certains tests

Étendre à l'anglais

- Suffixes -*ion*, -*ing*, -*ment*...
- Corpus *Wikipedia* anglais
- Utiliser la couche morphologique de WordNet (Felbaum, 1998)

Perspectives (2)

Comment caractériser la technicité ou la sous-spécification des voisins ?

- TF-IDF (distribution dans les articles)
- Nombre de sens (dictionnaire)
- Nombre de domaines (dictionnaire)
- Fréquence
- Annotation manuelle

Balvet, A., Barque, L., Condette, M. H., Haas, P., Huyghe, R., Marin, R., and Merlo, A. (2011). La ressource nomage. confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus. *Traitement Automatique des Langues*, 52(3) :129–152.

Benveniste, E. (1975). *Noms d'agent et noms d'action en indo-européen*. Maisonneuve, Paris.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker ? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Botha, J. and Blunsom, P. (2014). Compositional morphology for word representations and language modelling. In *International Conference on Machine Learning*, pages 1899–1907.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334) :183–186.

Dawes, E. (2003). La féminisation des titres et fonctions dans la francophonie : de la morphologie à l'idéologie. *Ethnologies*, 25(2) :195–213.

Dubois, J. (1962). *Étude sur la dérivation suffixale en français moderne et contemporain : essais d'interprétation des mouvements observés dans le domaine de la morphologie des mots construits*. Paris : Larousse.

Felbaum, C. (1998). *Wordnet, an Electronic Lexical Database for English*. Cambridge : MIT Press.

Fradin, B. (2014). La variante et le double. *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, pages 109–147.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3) :146–162.

Hathout, N. (2009). *Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie*. PhD thesis, Université Toulouse le Mirail-Toulouse II.

Huyghe, R. (2014). La sémantique des noms d'action : quelques repères. *Cahiers de lexicologie*.

Huyghe, R. and Tribout, D. (2015). Noms d'agents et noms d'instruments : le cas des déverbaux en-eur. *Langue française*, (1) :99–112.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. *arXiv preprint arXiv :1405.3515*.

Kintsch, W. (2001). Predication. *Cognitive science*, 25(2) :173–202.

Kisselew, M., Padó, S., Palmer, A., and Šnajder, J. (2015). Obtaining a better understanding of distributional models of german derivational morphology. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 58–63.

Kisselew, M., Rimell, L., Palmer, A., and Padó, S. (2016). Predicting the direction of derivation in english conversion. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 93–98.

Koontz-Garboden, A. (2007). *States, changes of state, and the Monotonicity Hypothesis*. PhD thesis, Stanford University.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Laca, B. (2001). Derivation. *Language Typology and Language Universals : An International Handbook*, 2 :1214–1227.

Lapesa, G., Kawaletz, L., Plag, I., Andreou, M., Kisselew, M., and Pado, S. (2017). Disambiguation of newly derived nominalizations in context : A distributional semantics approach.

Le Draoulec, A. and Péry-Woodley, M.-P. (2016). La femme de l'écrivain. `http://bling.hypotheses.org/1405`. Repéré le 20 avril 2017.

Lenoble-Pinson, M. (2008). Mettre au féminin les noms de métier : résistances culturelles et sociolinguistiques. *Le français aujourd'hui*, (4) :73–79.

Luong, T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113.

Martin, F. (2010). The semantics of eventive suffixes in french. *The Semantics of Nominalizations across Languages and Frameworks, Berlin, Mouton de Gruyter*, pages 109–141.

Mel'čuk, I. (2000). Un fou/une folle : un lexème ou deux ? *Lexique, syntaxe et sémantique. Mélanges offertes à Gaston Gross à l'occasion de son soixantième anniversaire*, pages 95–106.

Mikolov, T., Chan, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Padó, S., Palmer, A., Kisselew, M., and Šnajder, J. (2015). Measuring semantic content to assess asymmetry in derivation. In *Workshop on Advances in Distributional Semantics*.

Padó, S., Snajder, J., and Zeller, B. D. (2013). Derivational smoothing for syntactic distributional semantics. In *ACL (2)*, pages 731–735.

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony : A case study. *Linguistics*, 54(1) :149–188.

Plénat, M., Lignon, S., Serna, N., and Tanguy, L. (2002). La conjecture de pichon. *Corpus*, (1).

Riedl, M. and Biemann, C. (2016). Unsupervised compound splitting with distributional semantics rivals supervised methods. pages 617–622.

Roché, M. (2009). Pour une morphologie lexicale. *Mémoires de la Société de Linguistique de Paris*, 13(Nouvelle série n ° 17) :65–87.

Schulte Im Walde, S. (2006). Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2) :159–194.

Soricut, R. and Och, F. J. (2015). Unsupervised morphology induction using word embeddings. In *HLT-NAACL*, pages 1627–1637.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of artificial intelligence research*, 37 :141–188.

Varvara, R., Lapesa, G., and Padó, S. (2016). Quantifying regularity in morphological processes : An ongoing study on nominalization in german. *ESSLLI DSALT Workshop : Distributional Semantics and Semantic Theory*.

Verhoeven, B., Daelemans, W., and van Huyssteen, G. (2012). Classification of noun-noun compound semantics in dutch and afrikaans. In *Proceedings of the Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa*, pages 121–125.

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a man's wikipedia ? assessing gender inequality in an online encyclopedia. In *ICWSM*, pages 454–463.

Zeller, B., Šnajder, J., and Padó, S. (2013). Derivbase : Inducing and evaluating a derivational morphology resource for german. In *Proceedings of the 51st*

Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), volume 1, pages 1201–1211.

Zeller, B. D., Padó, S., and Snajder, J. (2014). Towards semantic validation of a derivational lexicon. In *COLING*, pages 1728–1739.