Confrontation of derivational processes and semantic categories in distributional semantic models

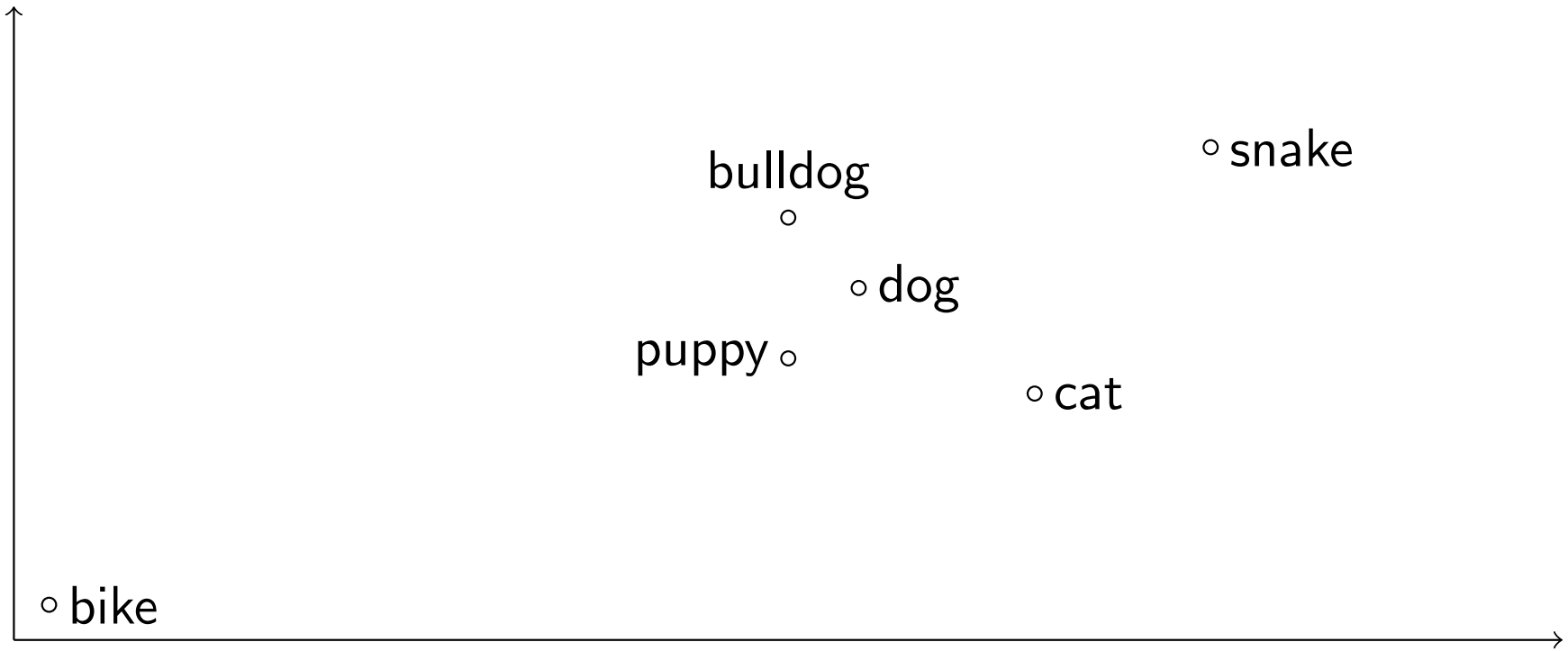
Marine Wauquier

LLF Morphology Workgroup

January 15th, 2021

Outline

- Rising popularity of Distributional Semantics
  - Spatial representation of meaning
  - Access to semantic properties



My objectives

- Challenges
  - Black box
  - Instability of representations
  - Class representation
- Powerful tool for semantic investigation
  - Flexible
  - Sensitive
  - Validation and exploration

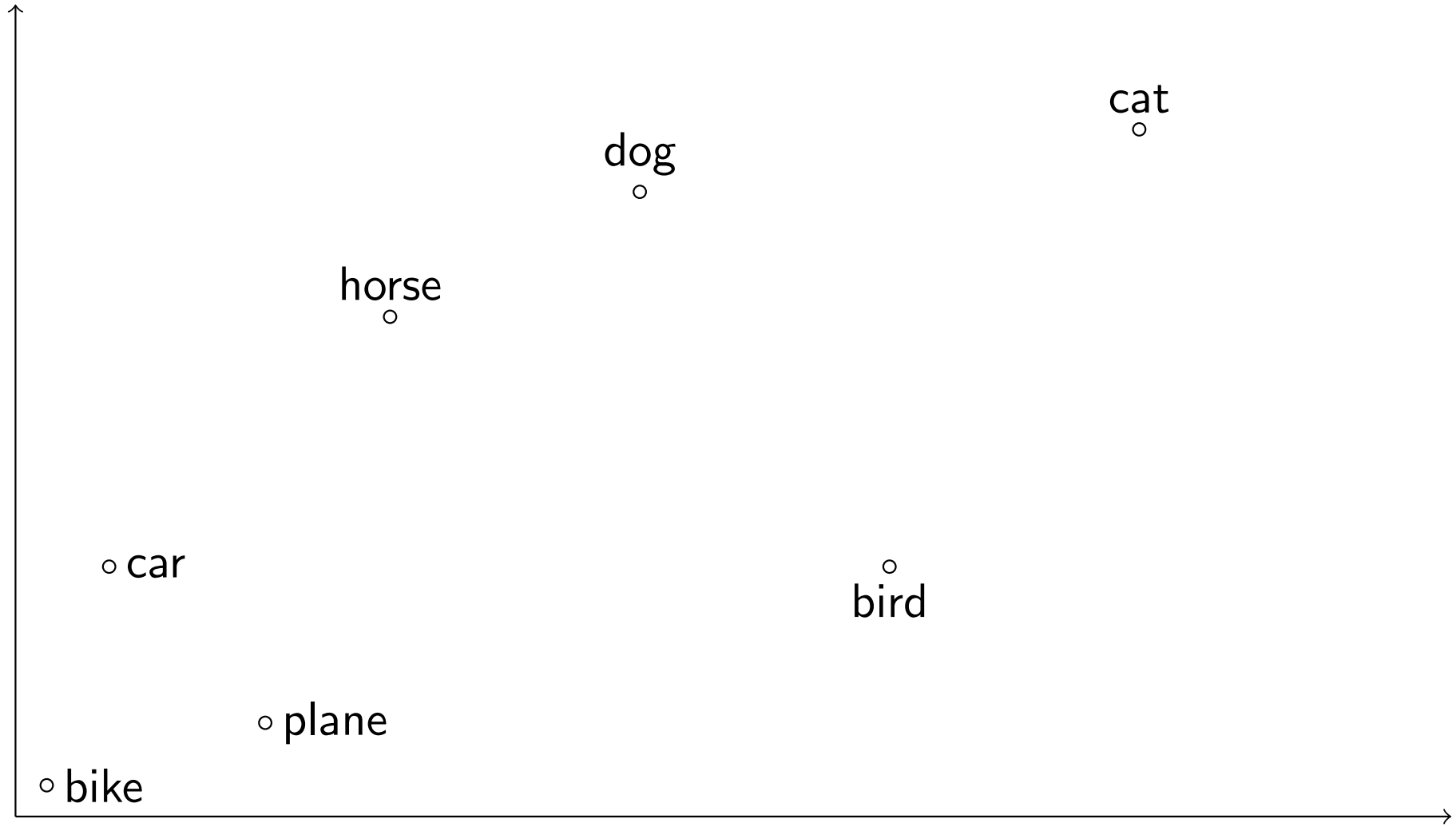
- Four main studies
  - Proximity within derivational families
  - Affix rivalry between -*euse* and -*rice*
  - Morphosemantic characterization of agent nouns
  - Discrimination of -*age*, -*ion* and -*ment* action nouns
- Contributions
  - Methodology fit for the semantic study of groups of words
  - New semantic insights on morphologically constructed nouns

Outline

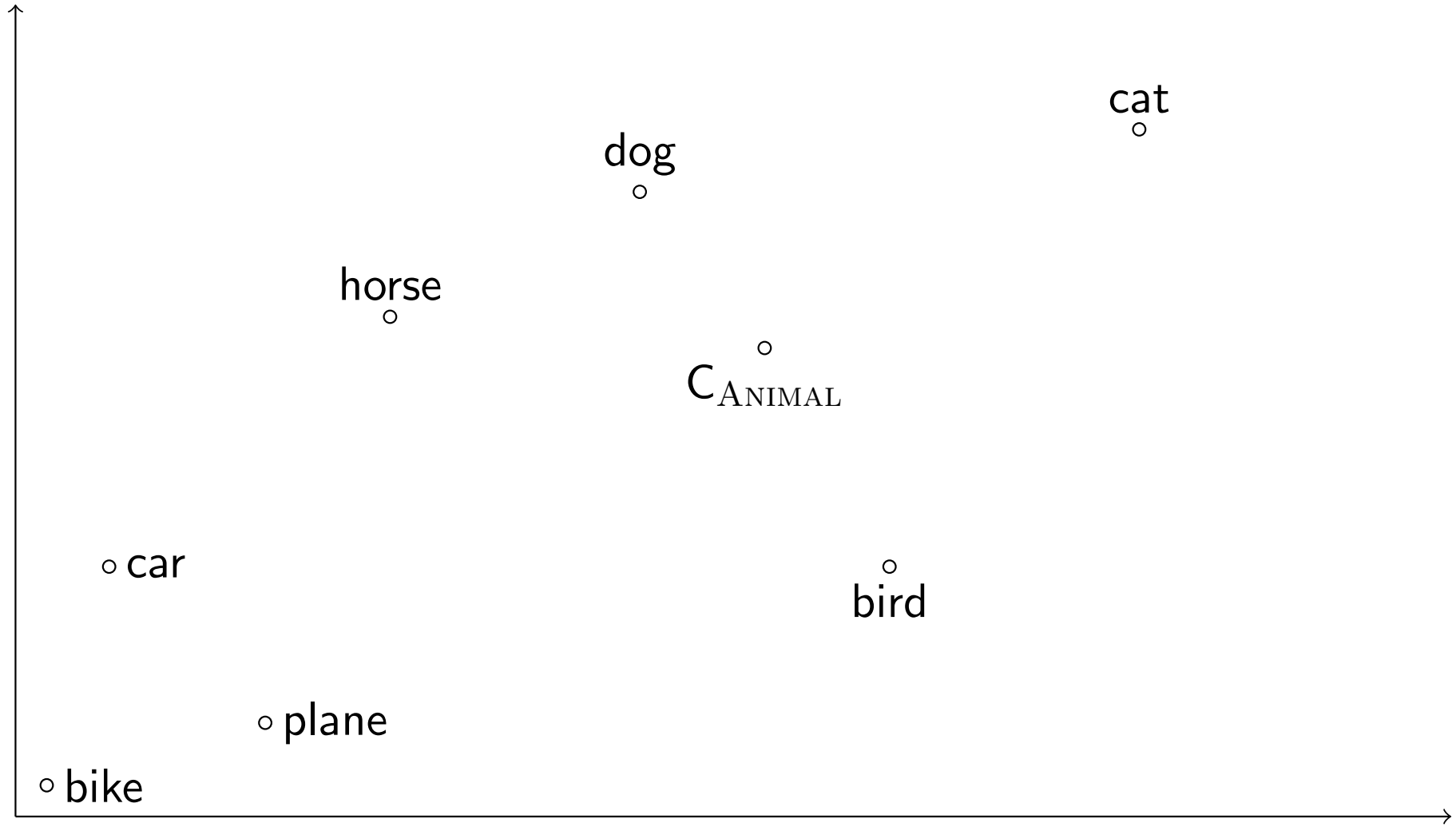
-*euse* and -*rice* differentiation

- Difference of connotation between rival suffixes -*euse* and -*rice*
  - -*euse* *(serveuse)* considered as more depreciative than -*rice* (*fondatrice*)
- Assessment of the semantic properties of -*euse* and -*rice* nouns at the scale of the group

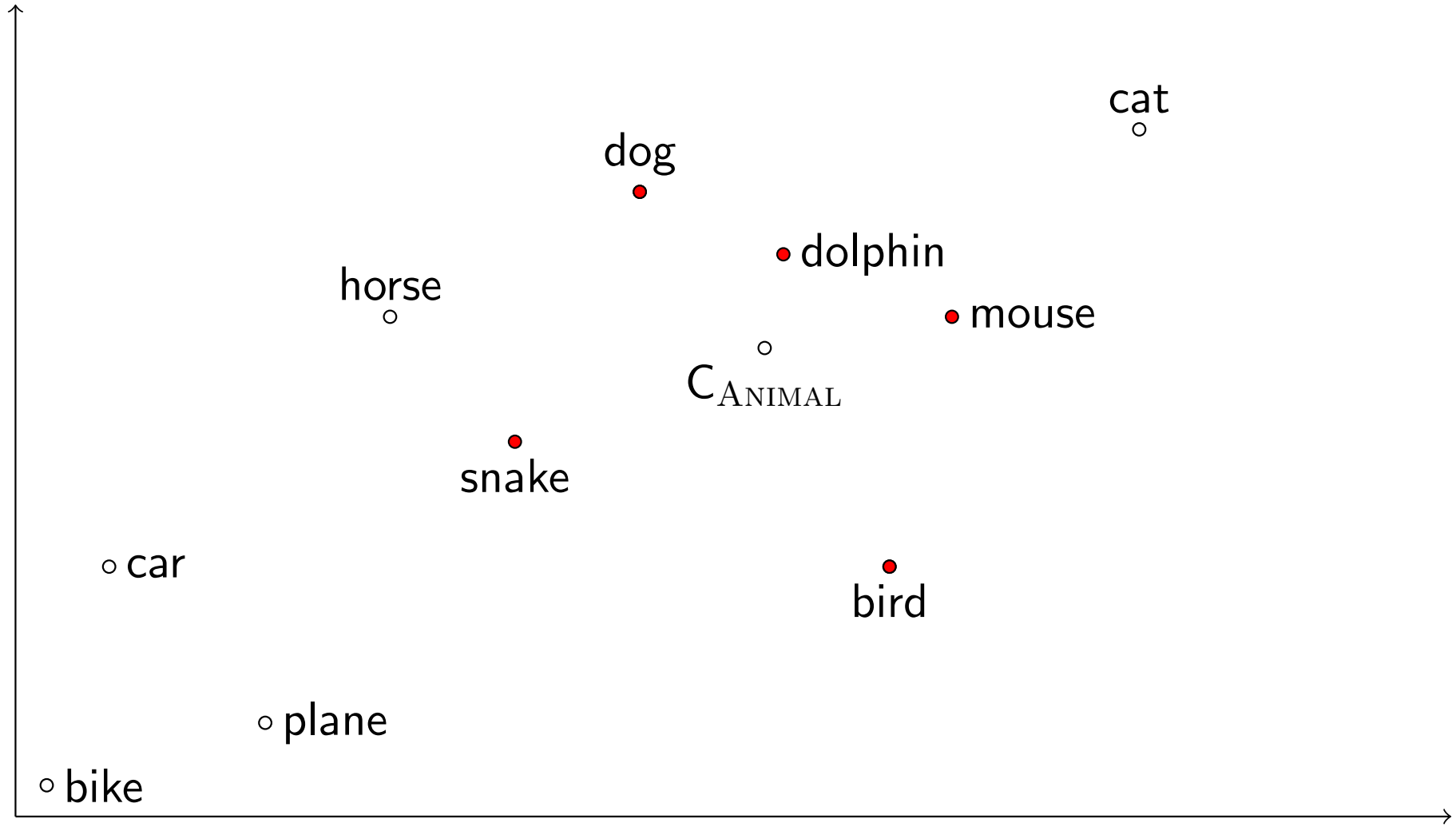
From words to class



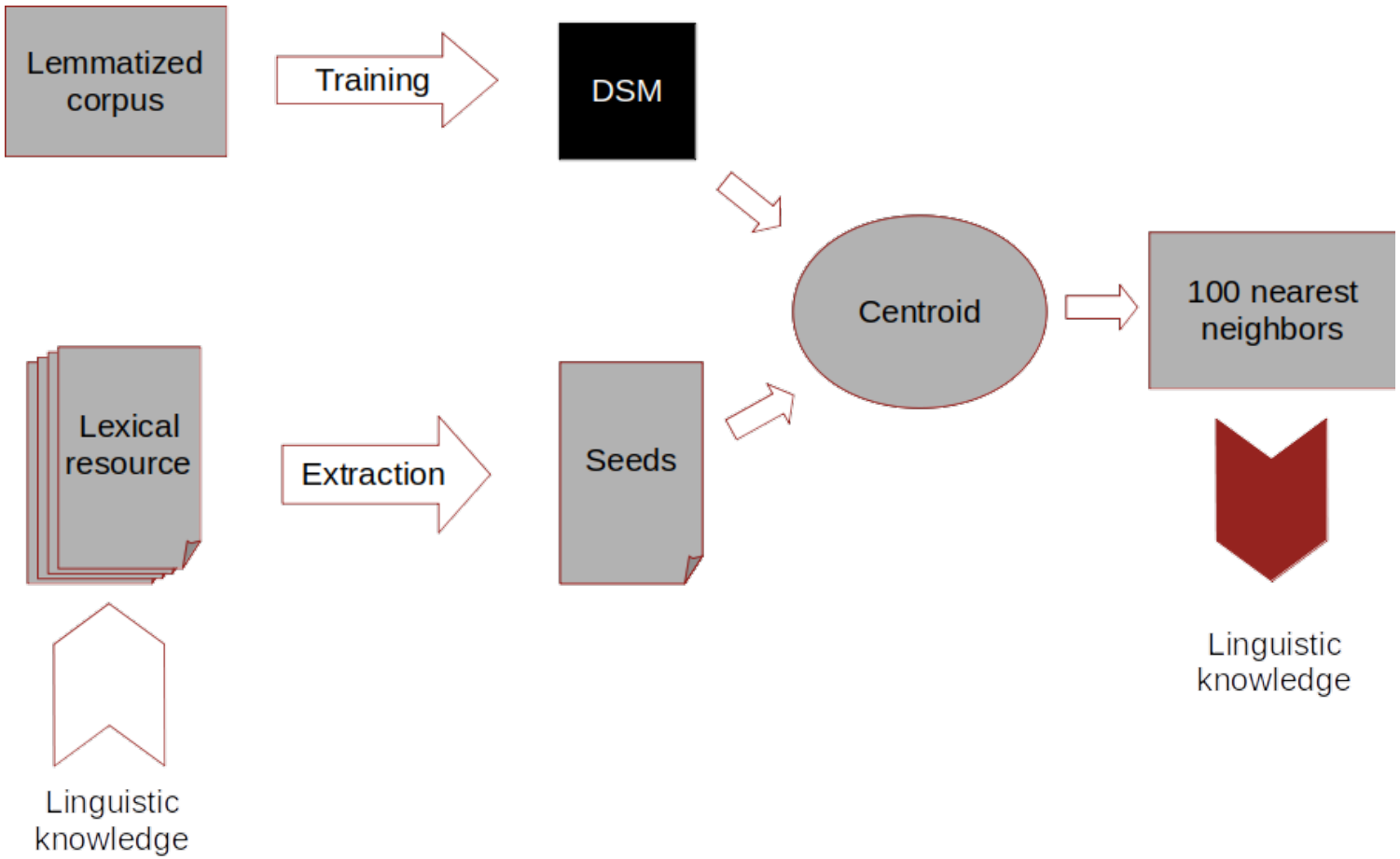
From words to class



From words to class



Processing chain



- Suffix -*euse*: depreciative feminine
  - agent nouns: *serveuse*, *stripteaseuse*, *boxeuse*, *coiffeuse*
  - instrument nouns: *cafetière*, *rôtissoire*, *essoreuse*, *tondeuse*
  - human and animal nouns: *midinette*, *gitane*, *tigresse*, *cochonne*, *chatte*
  - socio-cultural expectations: *bourriche*, *bigoudi*, *jupe-culotte*
- Suffix -*rice*: neutral feminine
  - agent nouns: *directrice*, *ingénieure*, *écrivaine*, *rédactrice*, *plasticienne*
  - instrument nouns: *grenailleuse*, *thermopile*

Limits of the analysis

- Number of items
  - 302 vectors for -*euse*
  - 77 vectors for -*rice*
- Frequency

	median	average
-*euse*	20	115
-*rice*	14	533

- Lemmatization
  - Over 80% of the 400.000 occurrences in -*euse* et -*rice* lemmatized with their masculine equivalent in -*eur* (*codétentrice* with *codétenteur*, *danseuse* with *danseur*) or -*eux* (*avantageuse* with *avantageux*)

Lexical class of agent nouns

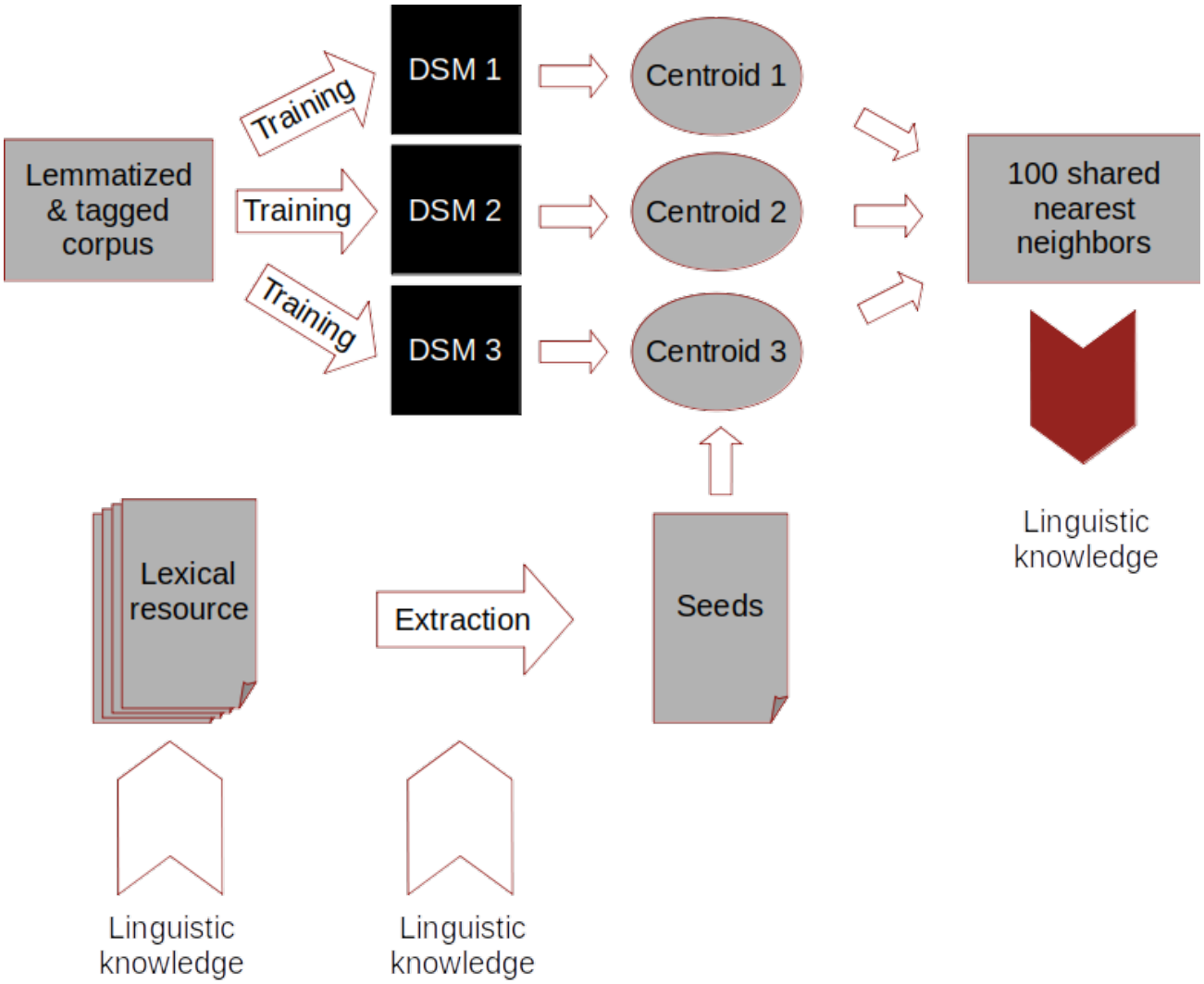
- Morphosemantic characterization of agent nouns
  - Morphological properties of agent nouns
  - Differentiation of *-ant*, *-aire*, *-eur*, *-ien*, *-ier* and *-iste*
  - Distinction between functional, occasional and behavioral agent nouns

Functional	Occasional	Behavioral
coiffeur	agresseur	bosseur
inspecteur	fondateur	charmeur

Fine-tuning of the method

- Fine-grained semantic issues
  - Increased control over data
    - Sufficient amount of monosemous agent nouns
    - Morphosyntactic annotation of the corpus to solve POS ambiguity
  - In-depth annotation of the 100 nearest neighbors
    - Morphological type, affix, POS and semantic type of the base

Processing chain



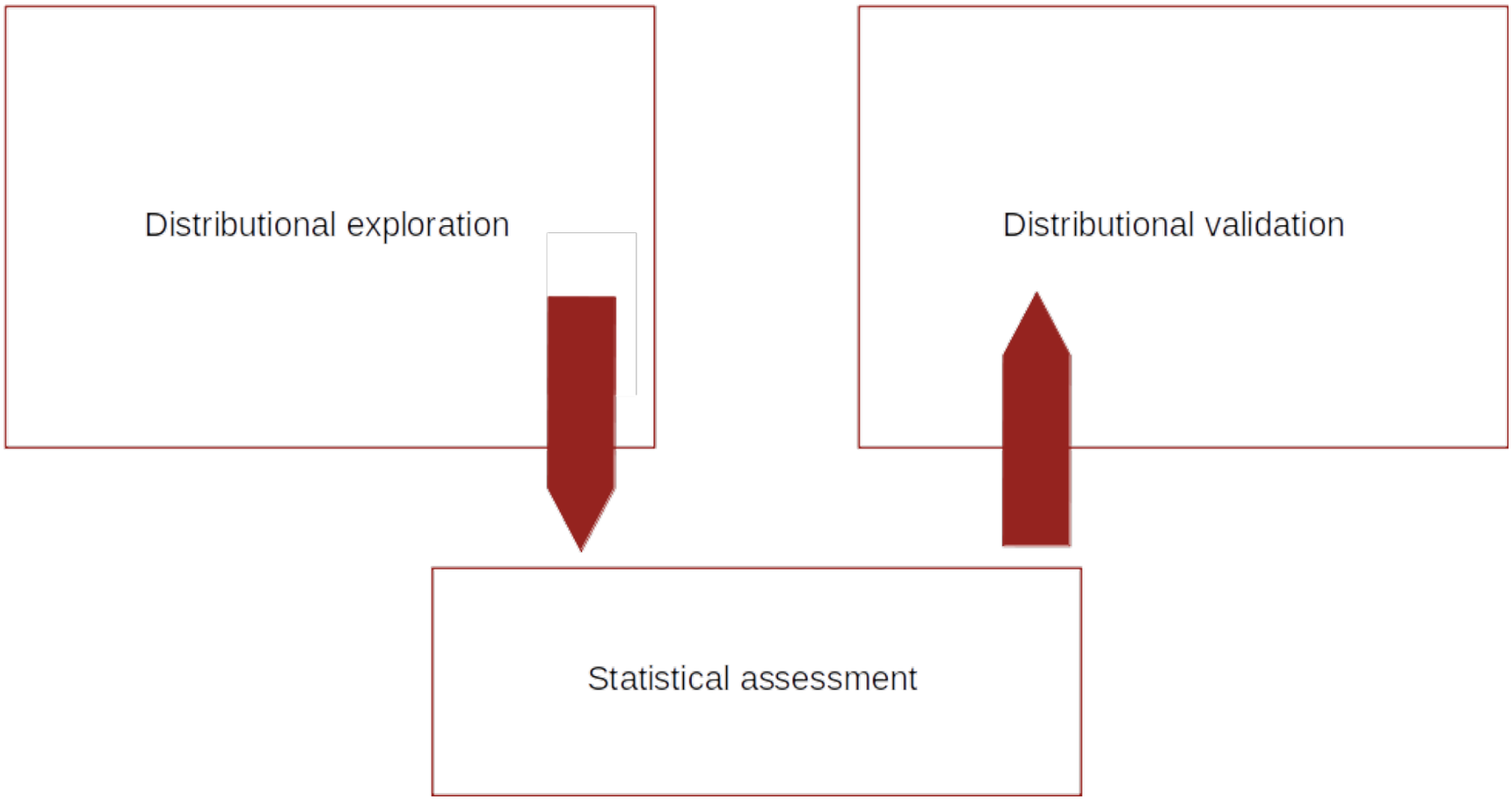
- Morphosemantic properties of agent nouns
  - Not necessarily morphologically constructed
- Differentiation of agentive suffixes
  - Similarity of -*eur* and -*ier* suffixes, and -*ien* and -*iste*
  - Importance of the semantic type of the base for the clustering of agent nouns
- Distinction between functional, occasional and behavioral agent nouns
  - Distinct areas of the DSMs
  - Morphosemantic properties
    - -*ier* and -*eur* suffixes for functional agent nouns
    - Adjectival bases for behavioral agent nouns
    - Action-denoting bases for occasional agent nouns

Outline

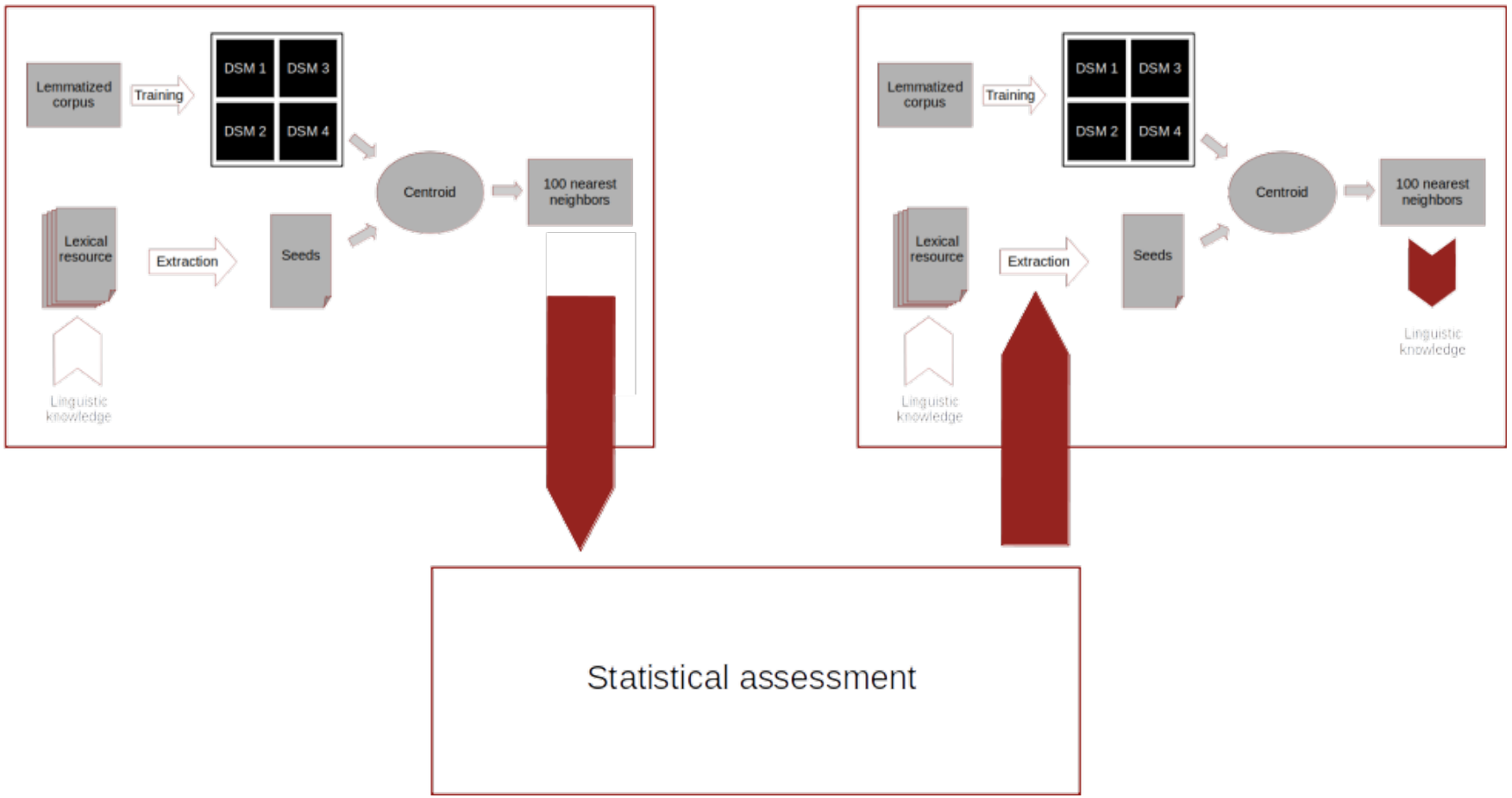
-*age*, -*ion* and -*ment* differentiation

- -*age*, -*ion* and -*ment* considered as rival affixes for action noun formation
  - Denotation of industrial processes for -*age* nouns
    - *usinage* and *meulage* vs *idéalisation* and *traitement*
- Three-step study
  - 1 - Distributional exploration of the three classes
  - 2 - Statistical assessment with corpus linguistics tools
  - 3 - Distributional validation

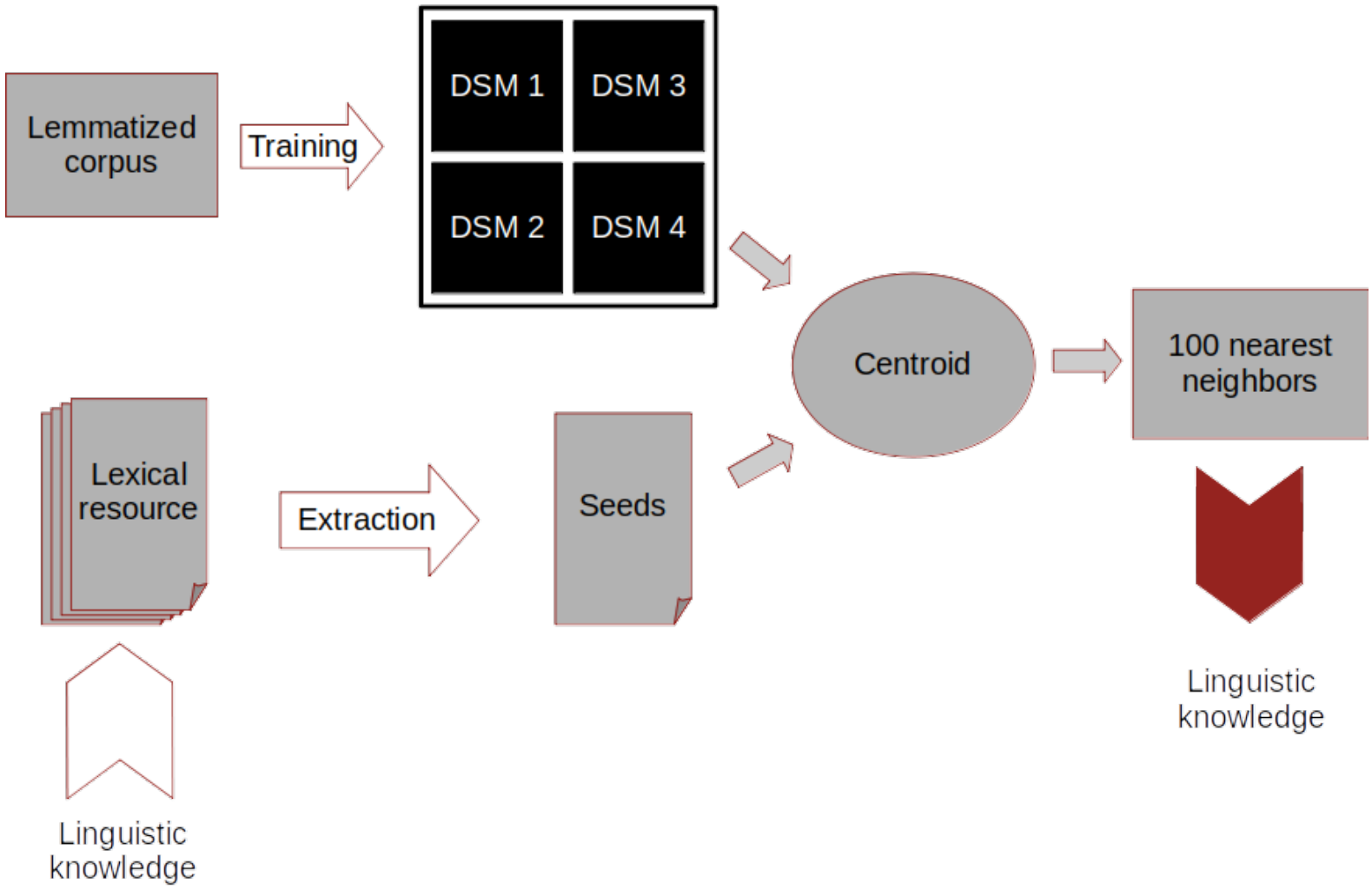
Processing chain (1)



Processing chain (2)



Distributional exploration



- Computation of *-age*, *-ion* and *-ment* centroids based on 629, 750 and 449 nouns respectively
  - Strong presence of seeds among the neighbors

-age	*-ion*	*-ment*
53%	45%	37%

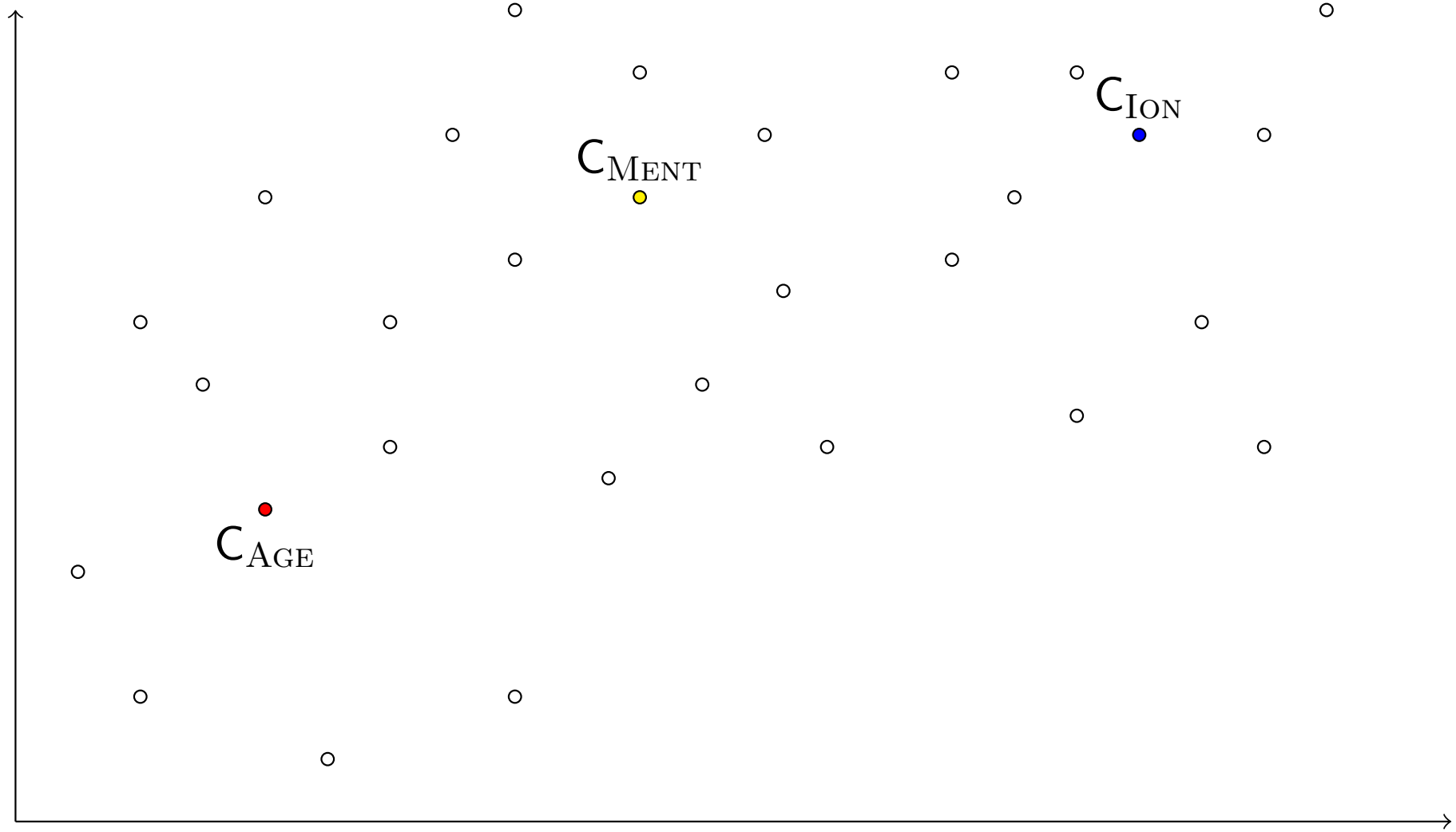
-eur	*-euse*	*-rice*
45%	22%	8%

  - Strong presence of the targeted suffix among the neighbors

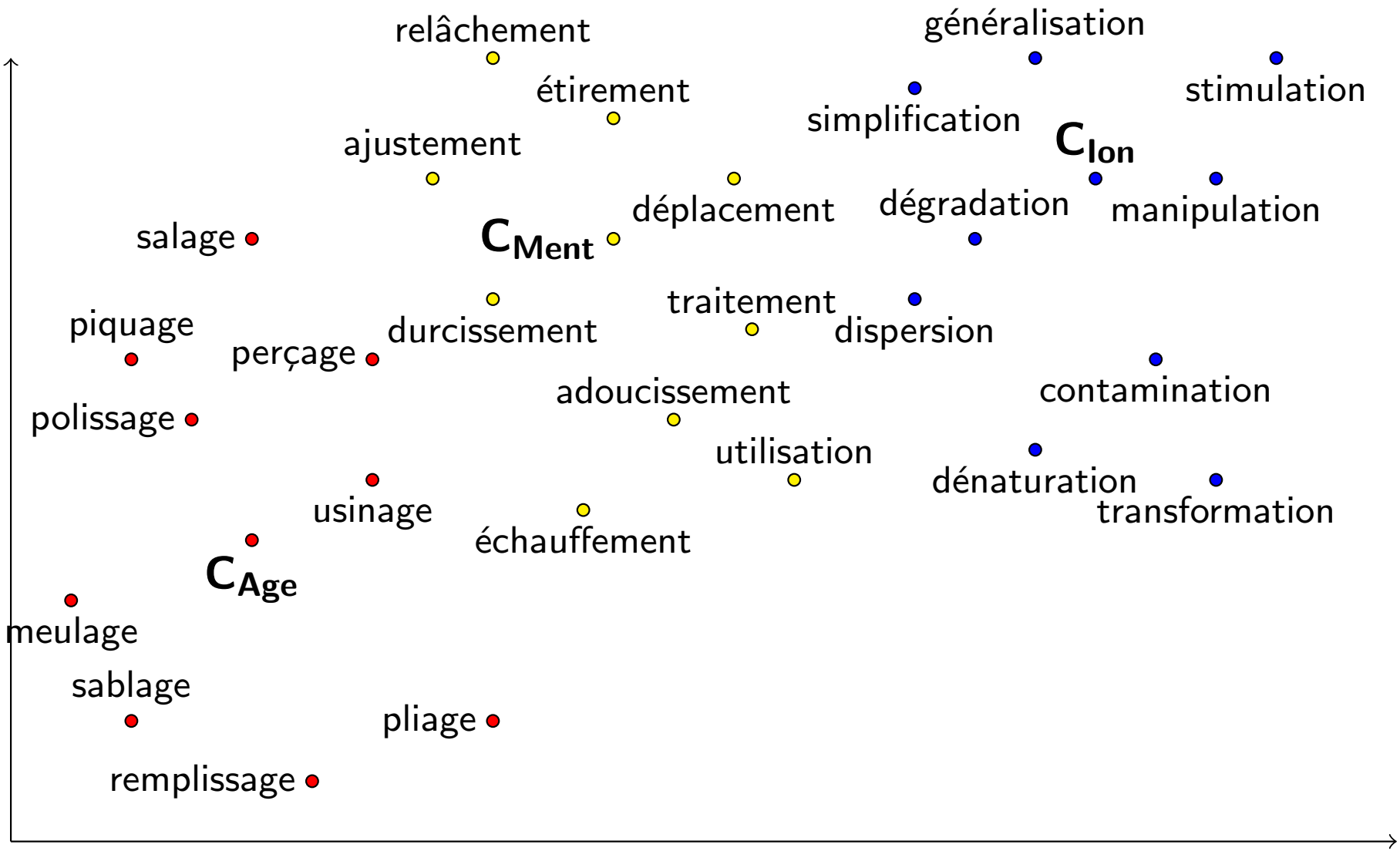
-age	*-ion*	*-ment*
82%	80%	73%

-eur	*-euse*	*-rice*
46%	27%	17%

Action nouns in vector space



Action nouns in vector space



Definition

A technical action noun is a noun **unfamiliar** to non-experts, denoting a **specific and complex action**, whose achievement and understanding require an **acquired skill** and which is specific to a particular domain. Technical action nouns typically, but not exclusively, belong to domains such as industry, agriculture and arts and crafts.

- Linguistic properties
  - T1 - Specialization
  - T2 - Obscurity
  - T3 - Univocity

Perceived technicality

- Annotation of 287 deverbal action nouns on a scale from 0 (non technical) to 5 (technical) by 3 speakers, normalized on a scale from 0 to 2
  - 47 *-age*, 47 *-ion*, 47 *-ification*, 47 *-isation* and 47 *-ment*
- Guidance
  - Agentivity
  - Complexity
  - Tools
  - Specificity
- Examples
  - *déclin* 0 (norm 0)
  - *danse* 1 (norm 0)
  - *inventorisation* 2 (norm 1)
  - *tondaison* 3 (norm 1)
  - *ciselure* 4 (norm 2)
  - *calandrage* 5 (norm 2)

Results of the main annotation

-age	*-ification*	*-ion*	*-isation*	*-ment*
0.7872	0.4894	0.2128	0.4894	0.1702

Annotation	*-age*	*-ification*	*-ion*	*-isation*	*-ment*
0	40.5%	57.4%	85.1%	57.4%	85%
1	40.5%	36.2%	8.5%	36.2%	13%
2	19%	6.4%	6.4%	6.4%	2%

Task difficulties

- Speaker intuition
  - *damage* 'tamping'
- Polysemy
  - *mirage* 'mirage', 'candling'
- Domains
  - *ancrage* 'integration', 'anchorage'
  - fabrication 'making'
- Annotators
  - Prior knowledge regarding the task objectives

- Moderate agreement
  - Annotation of 350 neighbors by 3 annotators
  - 0.247 on absolute values and 0.447 on normalized values (Fleiss kappa)
  - 0.45 on absolute values and 0.821 on normalized values after adjudication

Score	Abs	Norm
-1	1	1
0	0.390	0.831
1	0.269	0.689
2	0.463	0.894
3	0.439	na
4	0.674	na
5	0.468	na

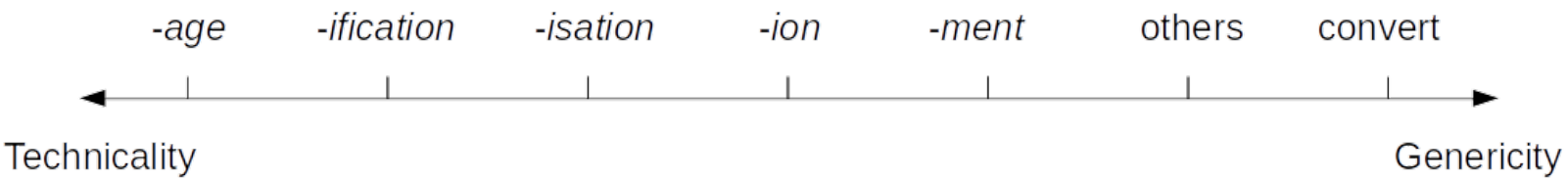
Property	Name	Description	Tech.
T1	RATIO_FREQR	Relative frequency ratio in *Wikipedia2018* and *LM10*	+
	NB_CAT_W18	Number of category markers in *Wikipedia2018*	–
	NB_DOM_T	Number of lexicographic domain markers in *TLFi*	–
	NB_DOM_G	Number of lexicographic domain markers in *GLAWI*	–
	LST	Presence or absence in *LexiTrans*	–
T2	PAGE_W18	Presence or absence of an entry in *Wikipedia2018*	+
T3	NB_SYN	Number of synonyms in *DES*	–
	NB_DEF_T	Number of definitions in *TLFi*	–
	NB_DEF_G	Number of definitions in *GLAWI*	–
	NSS	Presence or absence in *LexNSS*	–

Technicality criteria scores

- Annotation of 1828 deverbal action nouns

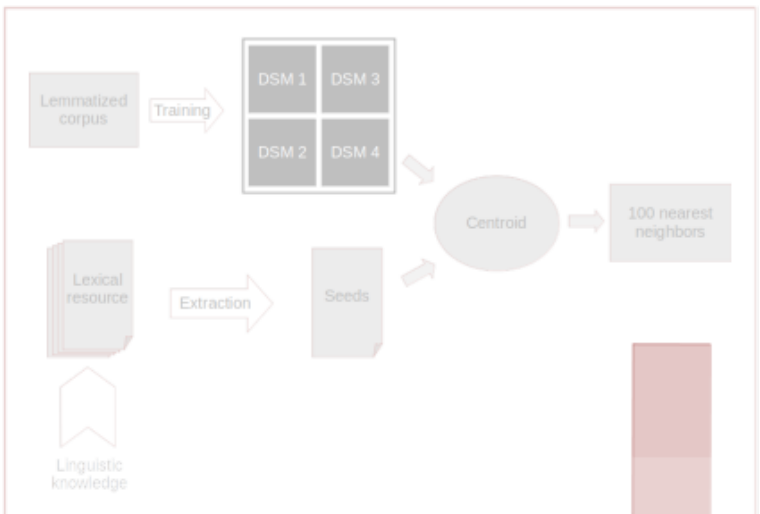
Criteria	*-age*	*-ification*	*-ion*	*-isation*	*-ment*
RATIO_FREQR (10^5)	4.81	1.66	3.02	1.14	3.05
NB_CAT_W18	0.78	1.04	0.90	1.02	0.42
NB_DOM_T	1.2	1.06	3.04	1.51	1.27
NB_DOM_G	0.65	0.51	1.04	0.60	0.46
LST (%)	0.2	2.68	9.14	8.51	2.23
PAGE_W18 (%)	48.97	57.14	72.42	68.09	34.08
NB_SYN	3.03	2.38	16.43	11.01	11.01
NB_DEF_T	2.52	2.22	6.66	3.51	4.34
NB_DEF_G	2.01	1.54	2.79	1.83	1.98
NSS (%)	0	0	2.54	0	1.11

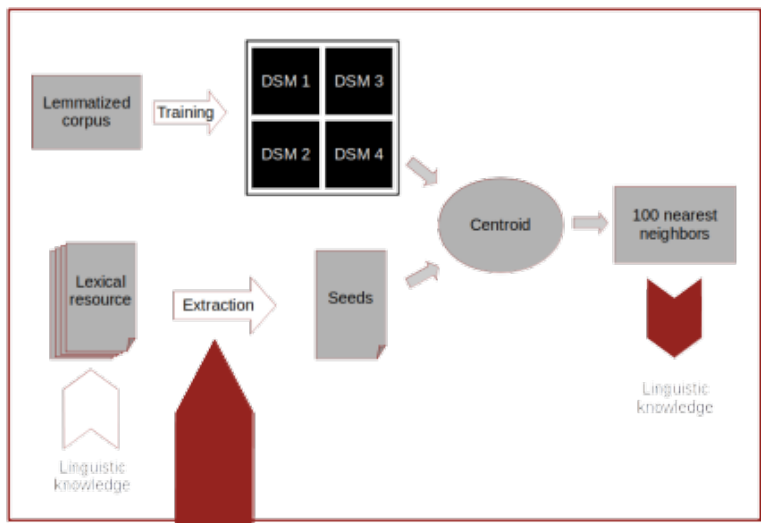
Action noun technicality



- Uneven discrimination of suffixes
  - Overall accuracy of 47%
  - 69% and 77% for convert and -*age* nouns
  - 31% and 40% for -*ment* and -*ion* nouns

Back to DSMs





Statistical assessment

- Technicality as a semantic feature captured by the DSMs
- Technicality of -*age*, -*ion* and -*ment* centroids

Technical nouns centroids

- Computation of centroids based on the perceived technicality scores
  - Tech0 - 198 non technical nouns (0)
  - Tech1 - 68 moderately technical nouns (1)
  - Tech2 - 21 technical nouns (2)

Criteria	Tech0	Tech1	Tech2
Ratio_freqr	0.81	3.28	8.14e+05
Nb_cat_w18	0.43	1.14	1.9
Nb_dom_t	5.6	3.73	2.75
Nb_dom_g	1.73	1.06	0.68
Lst (%)	32.5	11.8	3.51
Page_w18 (%)	82,5	86.2	90
Nb_syn	26.5	6.48	5.4
Nb_def_t	10.38	6.49	4.65
Nb_def_g	4.38	2.67	1.97
Nss (%)	5	0	0

Tehcnicality of -*age*, -*ion* and -*ment* centroids

- Computation of technicality criteria scores for the 100 nearest neighbors of -*age*, -*ion* and -*ment* centroids

	-*age*	-*ion*	-*ment*
Ratio_freqr	4.64e+05	1.60	1.06
Nb_cat_w18	1.19	0.78	0.37
Nb_dom_t	2.46	5.51	3.12
Nb_dom_g	0.9	1.66	0.92
Lst (%)	2	25	13
Page_w18 (%)	72	92	54
Nb_syn	4.11	19.02	19.49
Nb_def_t	4.54	9.5	6.75
Nb_def_g	2.41	3.82	3.03
Nss (%)	0	5	2

Comparison of morphological and technical centroids

- Agreement between morphological centroids and technical centroids
  - *-age* centroid closer to Tech2 centroid
  - *-ion* centroid closer to Tech0 centroid
- Assessment based on the number of shared neighbors

	-age	*-ion*	*-ment*
Tech0	2	66	12
Tech1	32	35	8
Tech2	53	5	1

- Strong semantic homogeneity of -*age*, -*ion* and -*ment* action nouns
  - Higher technicality of -*age*
- Assessment of technicality
  - Definition
  - Various measures and criteria
- Technicality as a semantic feature in the DSMs
- Use of DSMs at various stages of a complex study

Outline

Computational linguistics approach to shell nouns (Ho-Dac *et al.*, 2020)

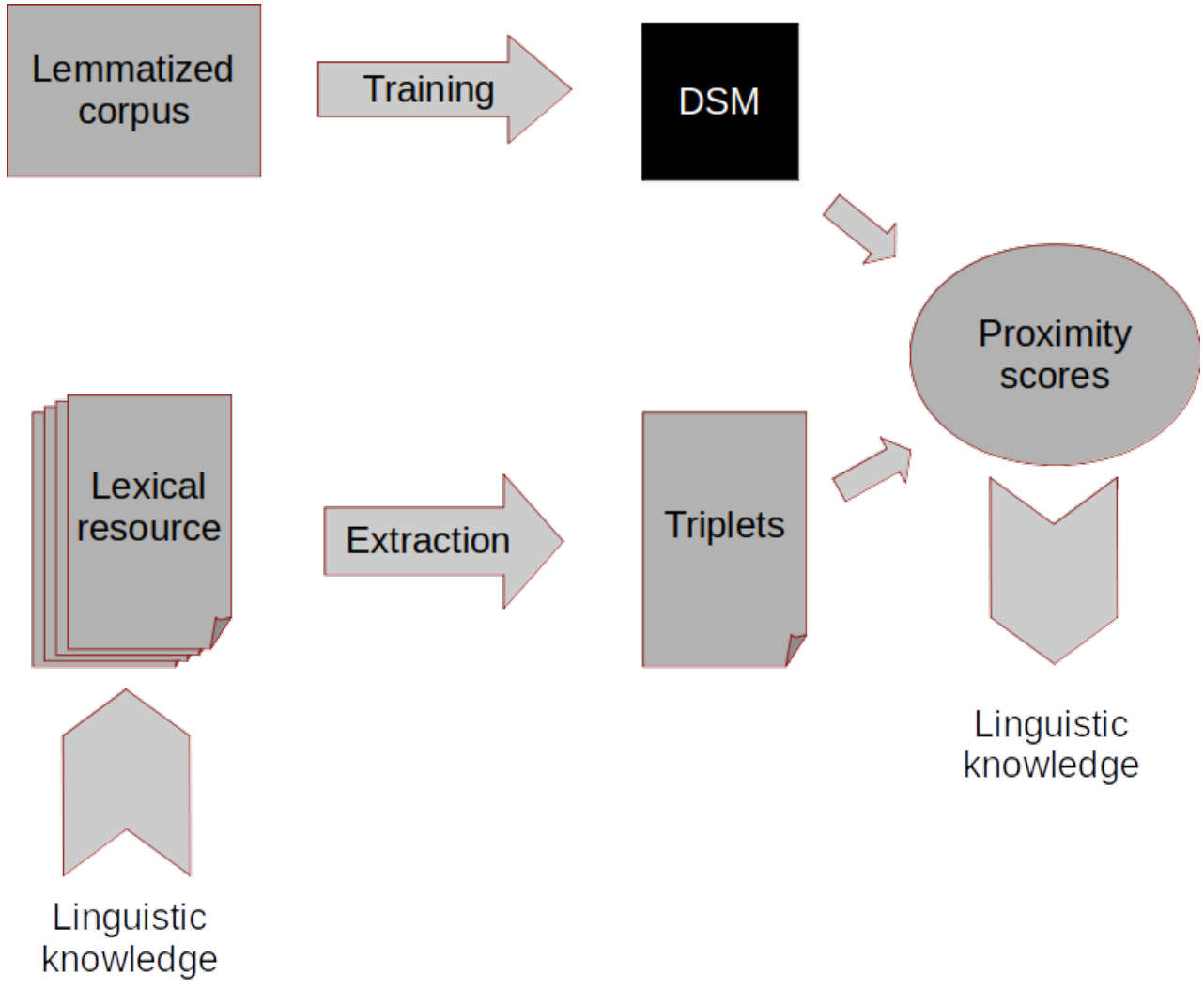
- Shell nouns
  - Discourse organization
  - Specific syntactic patterns (Riegel, 1996; Legallois, 2006)
  - General nouns, semantically underspecified (Halliday and Hasan, 1976)
- Identify shell nouns in corpora
  - Syntactic patterns
  - Distributional similarity
  - Two distinct corpora
    - wikiArt - descriptive and explanatory texts
    - wikiDisc - argumentative texts

- Centroid of 7 "canonical" shells nouns
  - *décision*, *fait*, *idée*, *possibilité*, *problème*, *question*, *raison*

wikiArt		**wikiDisc**	
question	0,78	raison	0.66
problème	0.72	question	0.62
fait	0.7	possibilité	0.61
idée	0.7	idée	0.6
problématique	0.69	objection	0.57
possibilité	0.68	difficulté	0.55
motivation	0.67	problème	0.54
raison	0.67	décision	0.53
conséquence	0.66	position	0.52
argument	0.66	hypothèse	0.52
décision	0.66	problématique	0.5
conclusion	0.64	notion	0.5
nécessité	0.63	conclusion	0.5
sujet	0.63	argument	0.5
notion	0.63	considération	0.48

- Neighbors of shell nouns centroids tend to be shell nouns (/50)
  - All neighbors are abstract nouns (except *donc* at rank 37 in wikiArt)
  - The majority of the 50 neighbors appear in "specificational" constructions (Legallois and Grea, 2006) (58% for wikiArt, 57,5% for wikiDisc)
- Corpus-specific shell nouns (42%)
  - wikiArt - *conséquence*, *sujet*, *principe*, *situation*, *réponse*, *discussion*,*affirmation*, *procédure*, *pratique*, *pertinence*
  - wikiDisc - *objection*, *position*, *obligation*, *option*, *responsabilité*,*remarque*, *justification*, *réticence*, *qualification*, *prétention*

Proximity scores



- Test the hypothesis of a higher proximity between the verb and its derived action noun
  - Annotation of the proximity score of 2585 {Verb, Agent noun, Action noun} triplets

Verb	Agent noun	Action noun	P(VbAg)	P(VbAc)	P(AgAc)
exécuter	*exécuteur*	*exécution*	0.373	**0.566**	0.375
itérer	*itérateur*	*itération*	0.503	**0.704**	0.343
détenir	*détenteur*	*détention*	**0.704**	0.160	0.106
réguler	*régulateur*	*régulation*	0.673	0.687	**0.754**

- Validation and quantification of the hypothesis

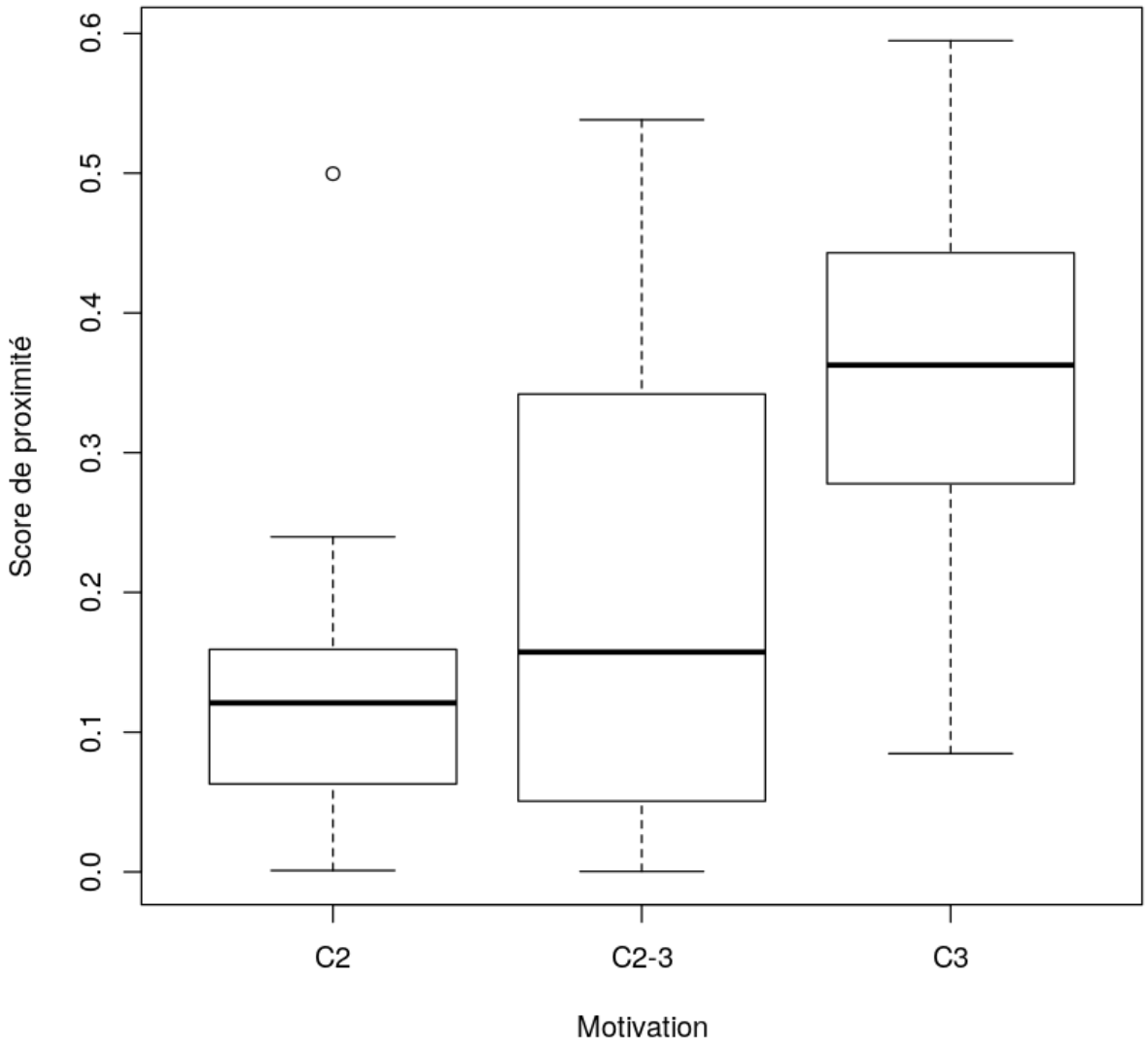
	VbAG	VbAc	AgAc
Average proximity score	0.253	0.400	0.289
Distribution of triplets	17%	59%	24%

Inherent limitations

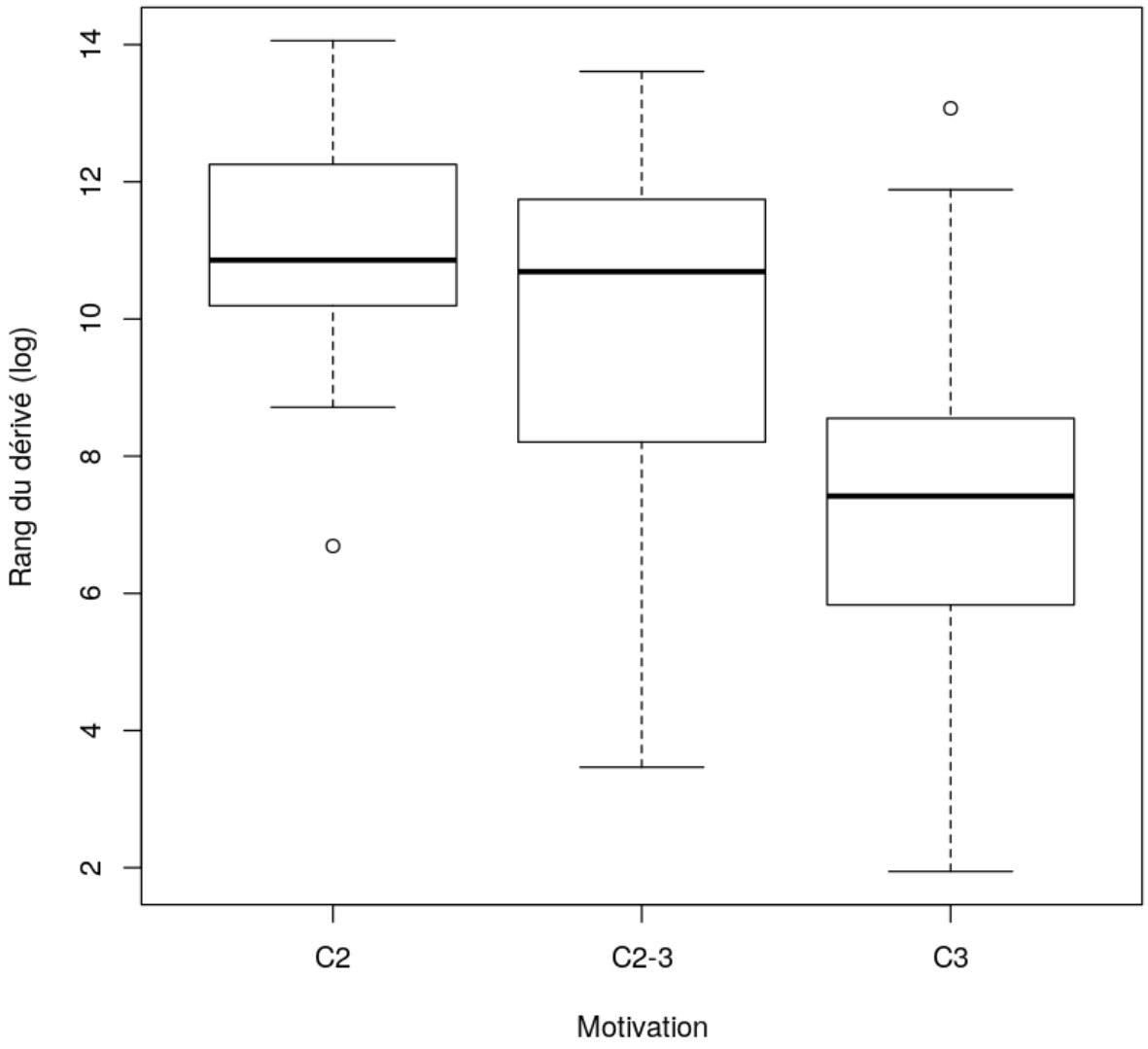
- Polysemy and homonymy
  - *porter* 'to carry' and *port* 'carrying', 'harbor'
  - *sprinter* 'to sprint' and *sprinteur* 'sprinter'
  - *goûter* 'to taste' and *goûter* 'snack'
- Lexicalization
  - *chauffer* 'to heat' and chauffeur 'car driver'
- Corpus
  - *franchir* 'to cross' and *franchisseur* 'pickup'
  - *mentir*, *menteur* and *menterie*

- Study of the motivation degree for pairs of nouns and verbs
  - *C0 - no link whatsoever : table 'table' and garer 'to park'*
  - *C1 - formal link : crevette 'shrimp' and crever 'to burst'*
  - C2 - demotivated link in synchrony : *peignoir* and *peigner*
  - C3 - motivated link in synchrony : *rasoir* 'razor' and *raser* 'to shave'
  - Additional C2-3 - ongoing demotivation : *fourrure* and *fourrer*
- Two-way assessment
  - Distributional proximity
  - Experimental assessment
- Hypothesis
  - Higher proximity score for C3 than C2-3 or C2
  - C3 items respectively closer within their corresponding neighborhoods than C2-3 or C3

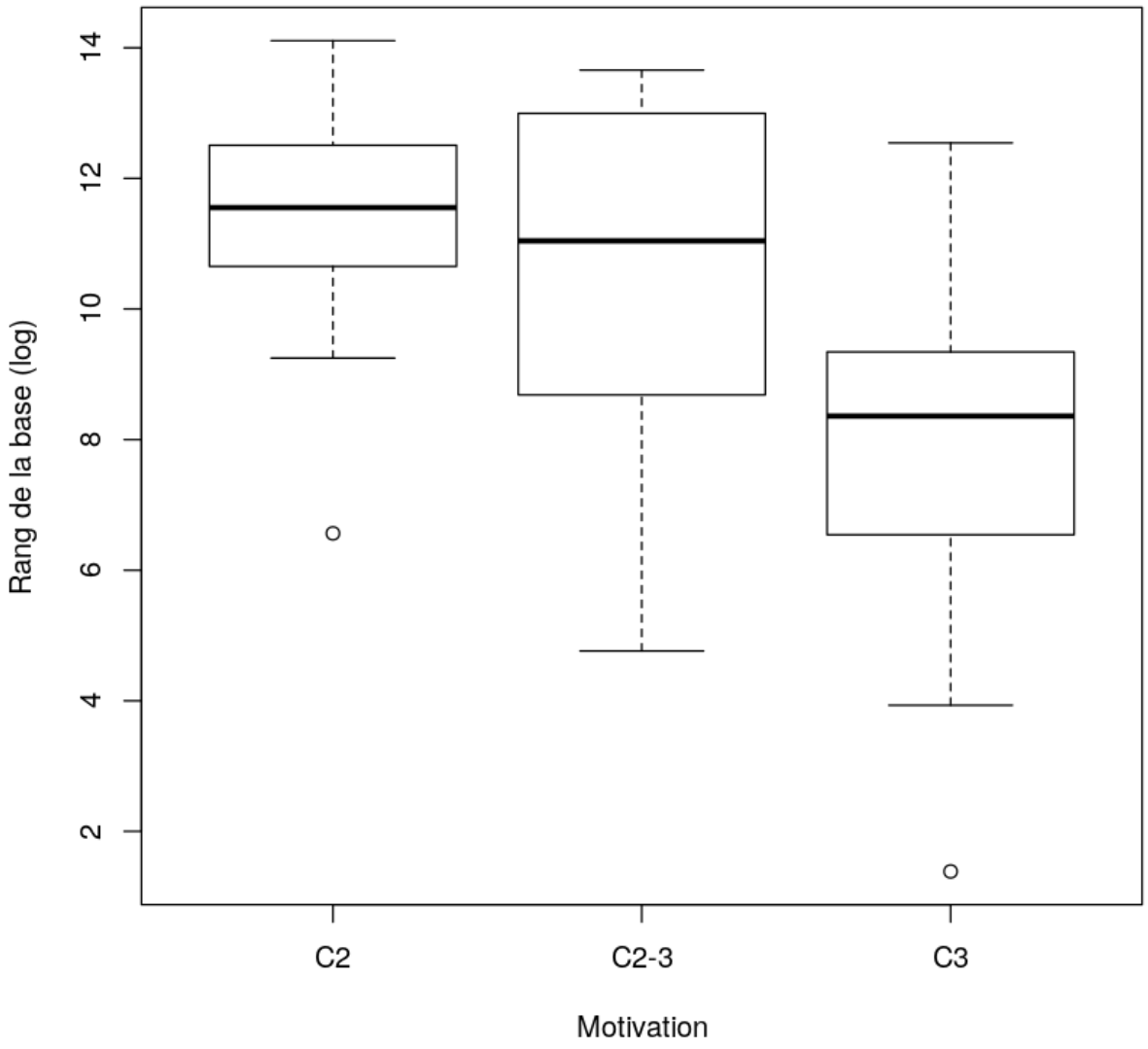
Proximity



Rank of the noun in the neighborhood of the verb



Rank of the verb in the neighborhood of the noun



Outline

Contributions

- Linguistic
  - Large scale validation of hypotheses
    - Higher proximity of verbs and action nouns
    - Depreciative connotation of *-euse*
  - New insights
    - Agent nouns
    - Technicality of action nouns
- Methodological
  - Centroid approach
  - Better understanding of DSMs potential

- Linguistic perspective
  - Telicity
  - Aspectual classification of action nouns
  - Lexicalization
- Methodological perspective
  - Contextual embeddings

- P(VbAc)

	Min	Average	Max
Agent noun	5	1344	261151
Verb	5	27992	1022519
Action noun	5	15014	422015

- P(AgVb)

	Min	Average	Max
Agent noun	5	4541	261151
Verb	5	39510	1842231
Action noun	5	15386	445155

- P(AgAc)

	Min	Average	Max
Agent noun	5	3799	325785
Verb	5	38227	1842231
Action noun	5	14898	404916

Distribution of frequency (1)

Figure: pVbAc

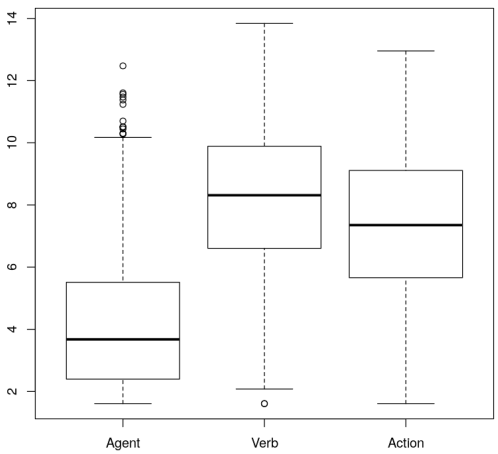


Figure: pAgVb

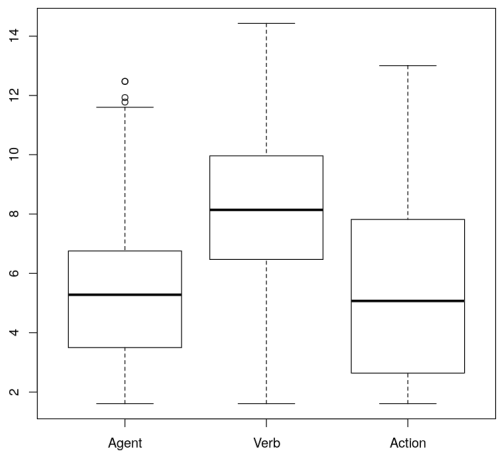
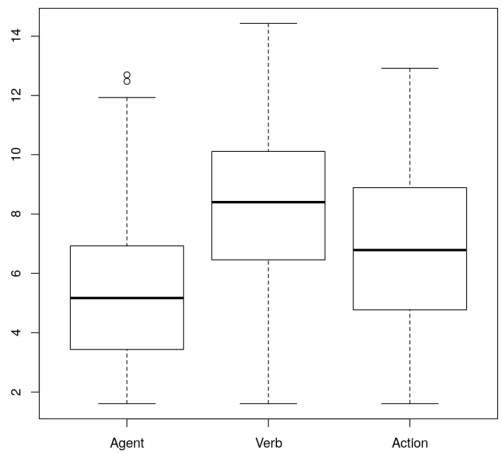
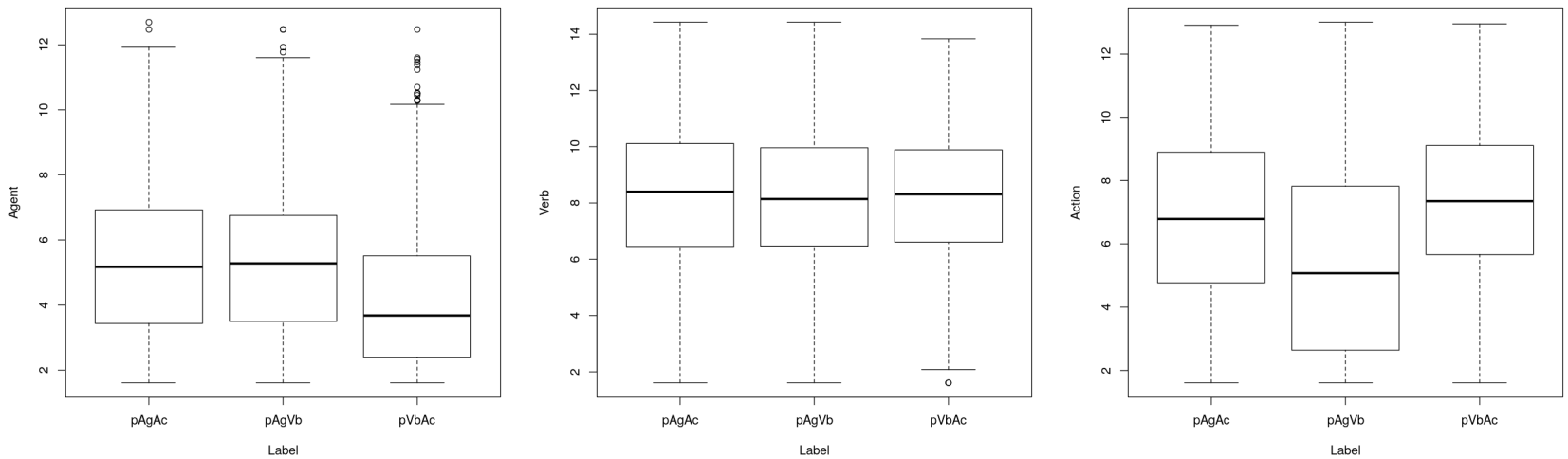


Figure: pAgAc



- Significant variation of frequency within the three groups

Distribution of frequency (2)



- Significant variation of agent and action nouns frequency across the three groups

Impact of the seeds on the neighbors analysis (/50)

	Affix.	Convert	Comp.	Complex	Simple	Extragram.	Indet.
Inclusive	31	4	2	5	4	0	4
Exclusive	28	4	3	7	4	0	4

	Adj	Noun	Verb
Inclusive	3	18	14
Exclusive	3	22	7

	Action	Object	Dom	Ppt	Instit	Obj Cog
Inclusive	17	11	2	3	2	0
Exclusive	12	12	3	3	2	0

	aire	ard	eur	ien	ier	iste	on
Inclusive	1	2	13	1	10	4	0
Exclusive	1	3	3	1	12	6	1

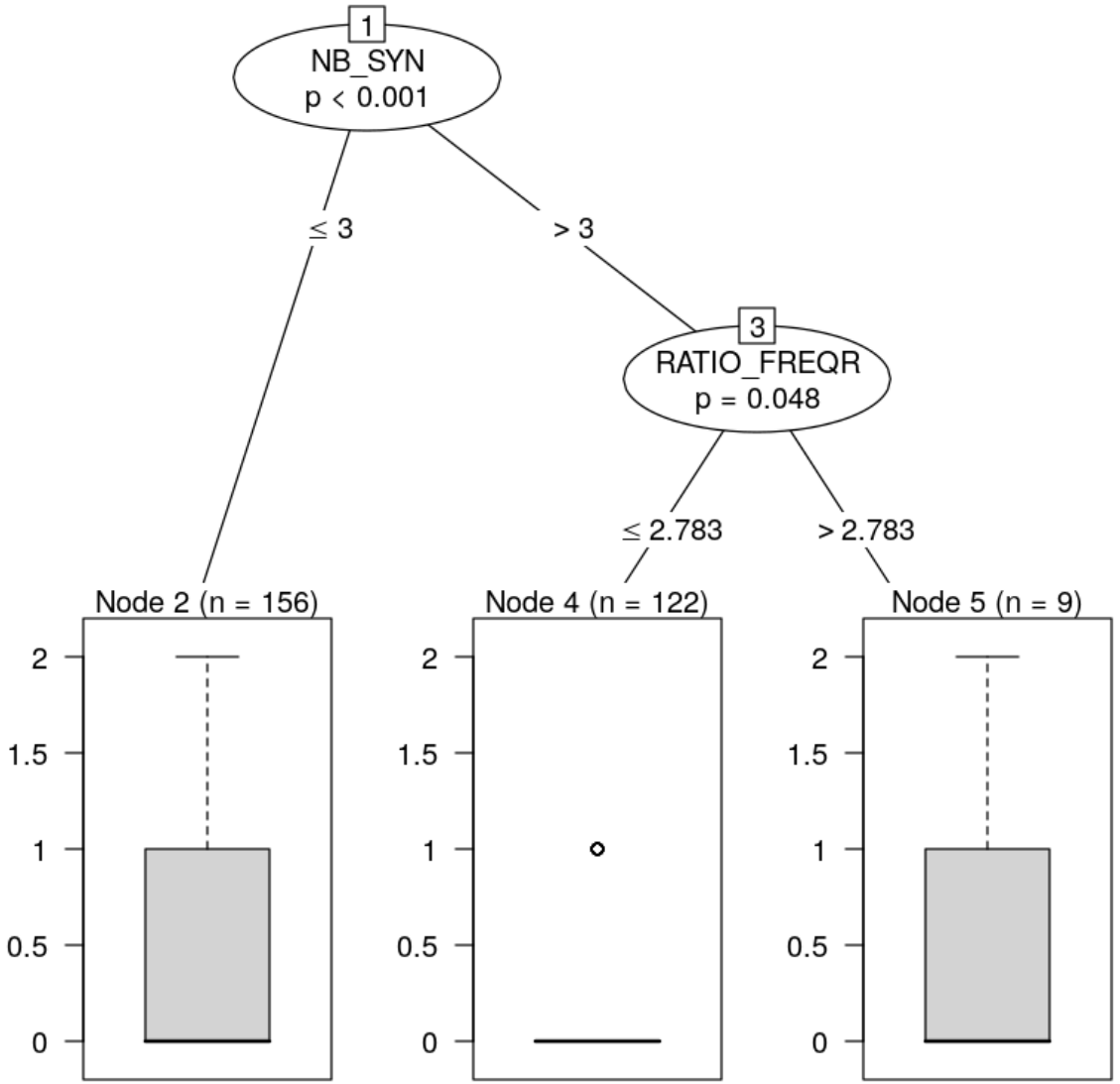
Impact of the seeds on the neighbors analysis

- On the 50 nearest neighbors
  - Never significant
- Variation of the analysis between 50 and 100 neighbors (non significant)
  - 40% of verbal base (50) vs. 52.1% (100)
  - 48.9% of action base (50) vs. 61.6% (100)
  - 31.4% of object base (50) vs. 21.9% (100)
- Seeds evenly distributed among the 100 nearest neighbors
- Variable number of seeds in the neighbors

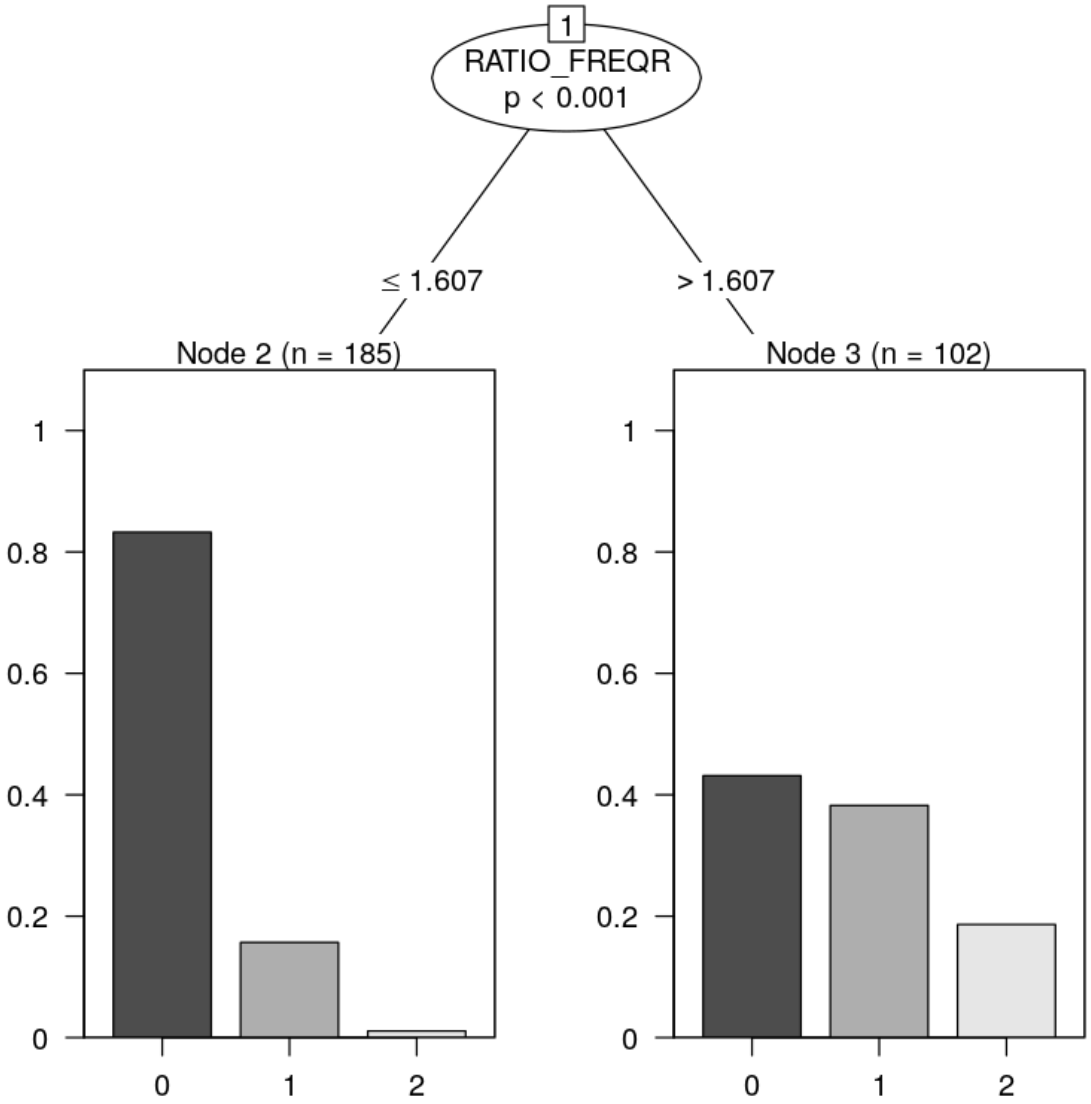
Linear regression for technicality scores

	Estimate	Std error	t value	Pr ($>$ |t|)
(Intercept)	0.392474	0.057011	6.884	3.77e-11
Ratio_freqr	0.017852	0.008345	2.139	0.03326
Page_w18	0.174830	0.075553	2.314	0.02139
Nb_syn	-0.005001	0.001725	-2.899	0.00404
Nb_def_t	-0.018720	0.006659	-2.811	0.00528

Conditional inference tree for technicality scores (1)



Conditional inference tree for technicality scores (2)



Decision tree for technicality scores