EVOLEX: un terrain pour éprouver les modèles et techniques TAL de mesures de proximité sémantique

Xavier de Boissezon[4] Lola Danet[4] Cécile Fabre[1] Jérôme Farinas[2] Bruno Gaume[1] Nabil Hatout[1] Lydia-Mai Ho-Dac[1] Mélanie Jucla[3] Patrice Péran[4] Bénédicte Pierrejean[1] Julien Pinquier[2] Ludovic Tanguy[1]

TMBI, 14.05.2018

(1) CLLE ERSS (2) IRIT CNRS - INPT - UPS - UT1 - UT2J (3) Octogone-Lordat (4) ToNIC Toulouse NeuroImaging Center UMR1214

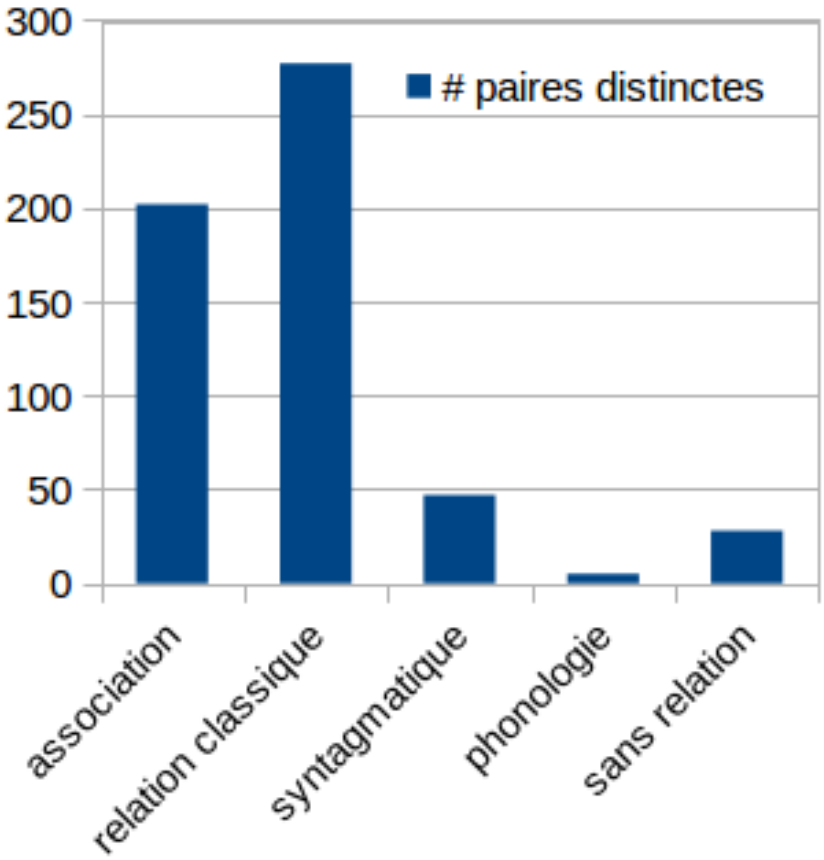
Paires de mots issues de la tâche de Génération

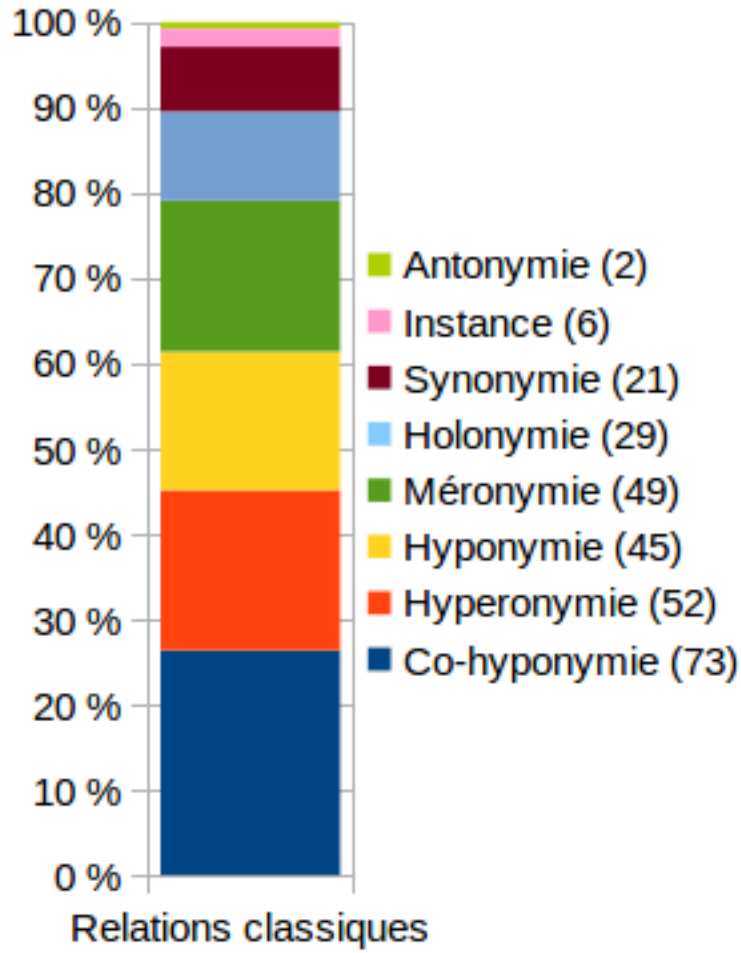
Données issues de V0 après correction des sorties du logiciel EvoLex

	A	AX	AY	AZ	BA	BB	BC	BD	
1	Stimulus	réponse	latence chrono	réponse	latence chrono	réponse	latence chrono	réponse	laten
39	Ortie	feuille	2,39	pissenlit	2,12	plante	2,05	piqure	
40	Osier	plante	2,87	panier	1,41	panier	1,89	panier	
41	Papillon	fleur	1,54	insecte	1,77	X	X	de nuit	
42	Parc	chien	1,48	jardin	1,46	jeux	1,25	cygne	
43	Perroquet	animal	1,52	oiseau	2,85	oiseau	1,68	bavard	
44	Pharmacie	médicament	1,39	médicament	1,96	médicament	1,74	sirop	
45	Pinceau	peinture	1,28	tableau	1,64	peinture	3,28	peintre	
46	Poil	barbe	1,92	rasoir	3,9	épilation	1,86	œuf	
47	Pôle	nord	2,73	X	X	nord	1,39	nord	
48	Rail	X	X	train	1,72	Faudel	2,18	train	
49	Raisin	figue	1,51	grappe	1,77	sec	1,19	vin	
50	Sac	récipient	2,06	portefeuille	3,51	à main	3,9	à ma	
51	Sapin	noël	1,05	noël	1,31	noël	0,97		
52	Serpent	reptile	1,48	sonnette	2,52	langue	1,26		
53	Siège	fauteuil	1,64	trône	2,5	fauteuil	1,48		
54	Sœur	frère	1,36	frère	1,54	frère	1,1		
55	Spaghettis	bolognaise	1,75	bolognaise	2,19	pâtes			
56	Terrier	lapin	0,96	renard	1,59	lapin			
57	Trou	vide	5,48	puits	2,85				
58	Truc	machin	1,08						
59	Varicelle	maladie							

Caractérisation des relations sémantiques entre mots

Double annotation manuelle puis adjudication





Caractérisation de la proximité sémantique entre mots

4 mesures : 2 ressources (usage vs. "norme") x 2 ordres

CORPUS collocations dans **FRWaC** (Baroni et al. 2009) : pages pages du domaine *.fr* (2 milliards de mots)

DICTIONNAIRE entrées et définitions dans le dictionnaire **TLF** *Trésor de la Langue Française* (Dendien and Pierrel 2003)

1er ORDRE la proximité entre A et B se mesure p.r. au fait qu'ils apparaissent dans un même contexte / que B apparait dans la définition de A

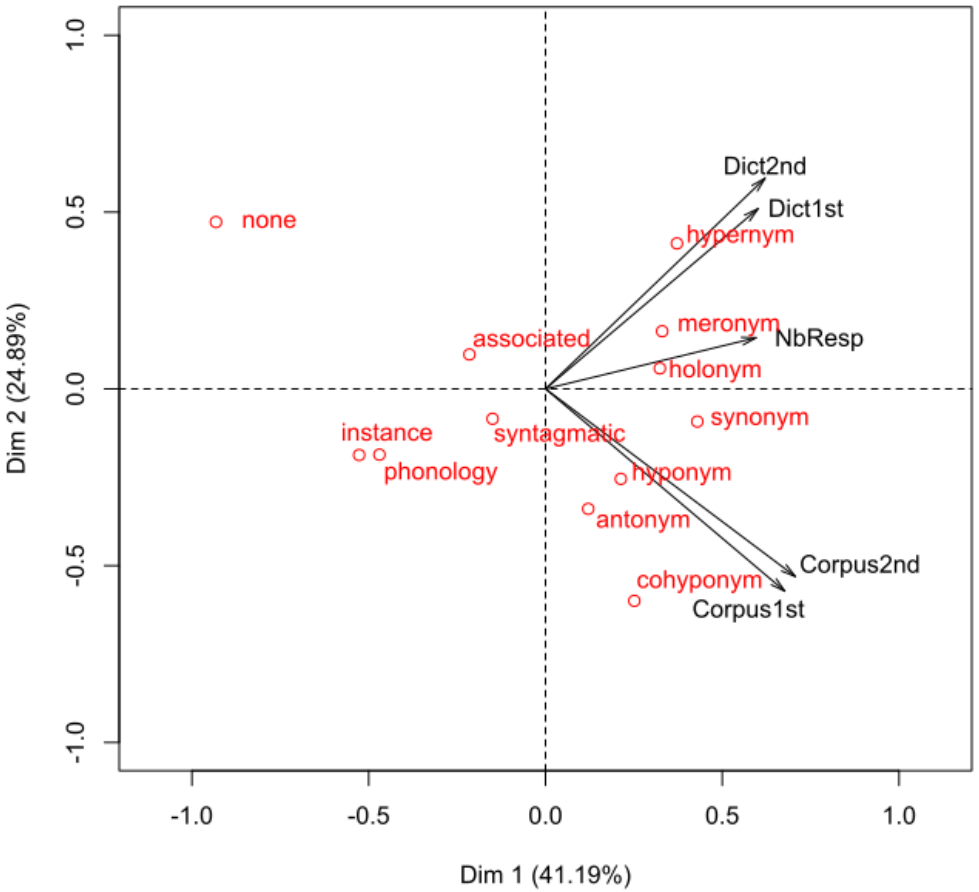
2nd ORDRE la proximité entre A et B se mesure p.r au fait qu'ils ont en commun des mots proches (voisins) de 1er ordre

Outils **Information Mutuelle** 1er ordre – corpus, **Word2Vec** (Mikolov et al. 2013) 2nd ordre – corpus, **Prox** (Gaume et al. 2016) 2nd ordre – dictionnaire

Caractérisation linguistiques des paires de mots

identifiant	(stimulus corrigé, r	SIMILARITE	ASSOC	SYNTAGMA	PHONO/MORPHO	?	POS	polysémie	N.jdm_r_associated.edge [0.34426541]	N.jdm_r_associated.prox [0.46134421]	N.jdm_r_syn.edge [0.08630895]	N.jdm_r_syn.prox [0.10639659]	FRWACDEP	FRWACCOOC
SRP-1	abricot,arbre	holonyme	x						90 : 0.71084663	0.00481111 : 1.27760392	0 : 0.00000000	0.00000000 : 0.00000000	0.307984140227	0.237063008347
SRP-2	abricot,caramel		x						0 : 0.00000000	0.00020092 : 0.05335488	0 : 0.00000000	0.00000000 : 0.00000000	0.716536036546	0.58595215267
SRP-3	abricot,confiture	holonyme		nden					110 : 0.86881255	0.00297627 : 0.79035695	0 : 0.00000000	0.00000000 : 0.00000000	0.699835570957	0.608131859695
SRP-4	abricot,fruit	hyperonyme							407 : 3.21460644	0.02385163 : 6.33386805	0 : 0.00000000	0.00000000 : 0.00000000	0.585114703217	0.498158440672
SRP-5	abricot,kiwi	cohypo							0 : 0.00000000	0.00039997 : 0.10621317	0 : 0.00000000	0.00000000 : 0.00000000	0.851678422504	0.601300971653
SRP-6	abricot,noyau	mero							193 : 1.52437111	0.00698082 : 1.85377657	0 : 0.00000000	0.00000000 : 0.00000000	0.355402325095	0.299773918085
SRP-7	abricot,orange	cohypo							110 : 0.86881255	0.00566499 : 1.50435418	0 : 0.00000000	0.00000000 : 0.00000000	0.625514479621	0.501379154017
SRP-8	abricot,pêche	cohypo							140 : 1.10576143	0.01580459 : 4.19695374	0 : 0.00000000	0.00000000 : 0.00000000	0.388459395455	0.304255311713
SRP-8	abricot,pêche	cohypo							140 : 1.10576143	0.01580459 : 4.19695374	0 : 0.00000000	0.00000000 : 0.00000000	0.388459395455	0.304255311713
SRP-9	ampoule,ampère		x					x	0 : 0.00000000	0.00014854 : 0.03944522	0 : 0.00000000	0.00000000 : 0.00000000	0.469875706529	0.282082027716
SRP-16	ampoule,baïonnette													
SRP-10	ampoule,brûlure		x					x	20 : 0.15796592	0.00088016 : 0.23372899	55.0 : 5.72600053	0.01274641 : 2.91072580	0.423211278474	0.311680853138
SRP-11	ampoule,électricité		x					x	139 : 1.09786313	0.00890664 : 2.36518353	0 : 0.00000000	0.00024113 : 0.05506361	0.333496831652	0.314671688215
SRP-12	ampoule,filament	mero						x	123 : 0.97149040	0.00376784 : 1.00056061	0 : 0.00000000	0.00009200 : 0.02100880	0.472884948322	0.376733679556
SRP-13	ampoule,lampe	holonyme							193 : 1.52437111	0.01416701 : 3.76208972	81.0 : 8.43283715	0.01651294 : 3.77083748	0.705805620646	0.701339379011
SRP-14	ampoule,lumière		x					x	400 : 3.15931837	0.01795581 : 4.76821632	0 : 0.00000000	0.00414214 : 0.94588467	0.366875715353	0.284478732093
SRP-15	ampoule,pied	holonyme						x	139 : 1.09786313	0.00397904 : 1.05664537	0 : 0.00000000	0.00000000 : 0.00000000	0.251624256718	0.234776410
SRP-17	animal,bestial						adj		-r_abs	-r_abs	-r_abs	-r_abs	0.0	0.0
SRP-18	animal,bestiole	syn							54 : 0.42650798	0.00005179 : 0.01375298	71.5 : 7.44380070	0.02141655 : 4.89060878	0.608287387978	0.3988
SRP-19	animal,chat	hypo							612 : 4.83375710	0.00042390 : 0.11256785	0 : 0.00000000	0.00001859 : 0.00424515	0.569137592677	0.5
SRP-20	animal,chien	hypo							672 : 5.30765486	0.00042859 : 0.11381329	0 : 0.00000000	0.00010261 : 0.02343166	0.60412835526	0
SRP-21	animal,compagnie			nden					55 : 0.43440628	0.00003196 : 0.00848707	0 : 0.00000000	0.00000000 : 0.00000000	0.190681582205	
SRP-22	animal,fouine	hypo							52 : 0.41071139	0.00007716 : 0.02049006	0 : 0.00000000	0.00000000 : 0.00000000	0.515076620	
SRP-23	animal,humain	hypo							0 : 0.00000000	0.00014104 : 0.03745357	0 : 0.00000000	0.00039320 : 0.08978978	0.451234	
SRP-24	animal,lion	hypo							361 : 2.85128483	0.00025391 : 0.06742652	0 : 0.00000000	0.00003101 : 0.00708134	0.455	
SRP-25	animal,mignon						adj		0 : 0.00000000	0.00001578 : 0.00419042	0 : 0.00000000	0.00012050 : 0.02751696	0.5	
SRP-26	animal,mouton	hypo							23 : 0.18166081	0.00011265 : 0.02991453	0 : 0.00000000	0.00000000 : 0.00000000	0.5	
SRP-27	animal,oiseau	hypo							61 : 0.48179605	0.00092663 : 0.24606923	0 : 0.00000000	0.00037876 : 0.08648		
SRP-28	animal,végétal	cohypo							54 : 0.42650798	0.00006368 : 0.01691040	0 : 0.00000000	0.00015060 : 0.03		
SRP-29	aube,aurore	syn						x	100 : 0.78982959	0.00552794 : 1.46796016	78.0 : 8.12050985	0.03		
SRP-30	aube,communion		x					x	0 : 0.00000000	0.00032927 : 0.08743858	0 : 0.00000000			
SRP-31	aube,crépuscule	antonyme						x	40 : 0.31593184	0.00293856 : 0.78034285	0 : 0.00000000			
SRP-40	aube,Hiroshima								0 : 0.00000000	0.00000000 : 0.00000000				
SRP-32	aube,jour	holonyme						x	20 : 0.15796592					

Corrélation entre les caractérisations (ACP)

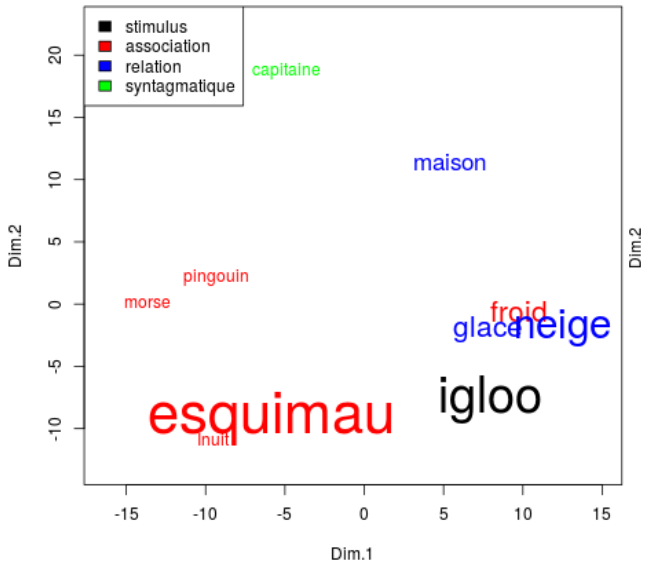


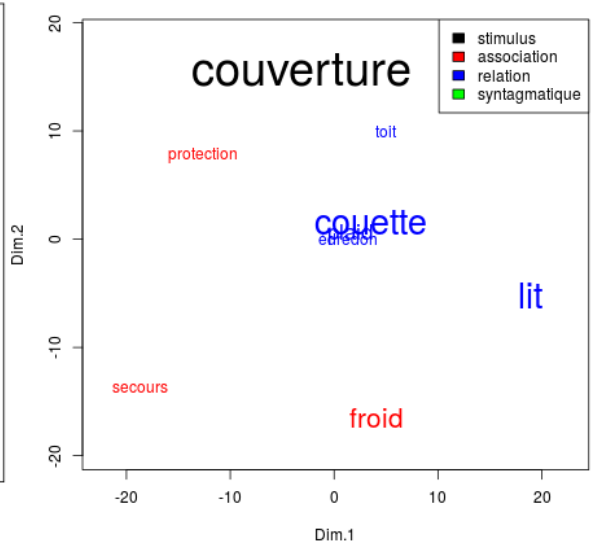
- Toutes les mesures sont positivement corrélées aux relations classiques
- Toutes les mesures montrent les scores les plus faibles pour les paires sémantiquement éloignées (sans relation, instance, phono)
- Les 2 Ressources / Méthodes évaluent des aspects différents du lexique (dictionnaire – hyperonymie ; corpus – cohyponymie)
- Association et Syntagmatique au centre (non corrélées, non caractérisées)

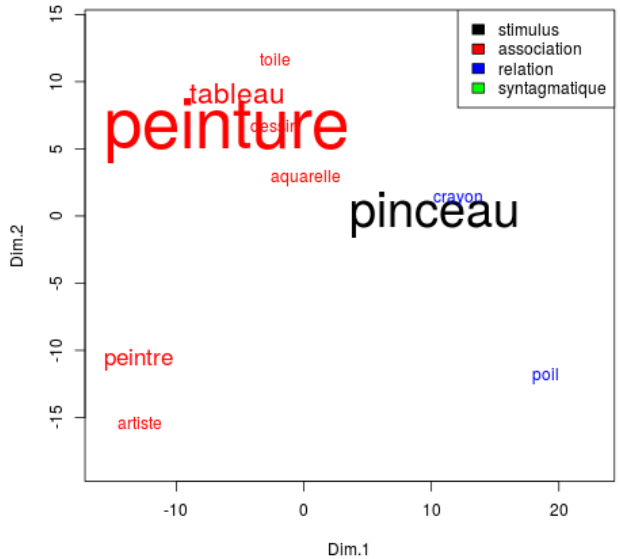
Au delà de la caractérisation des paires de mots...

Identifier des groupes (catégories ? profils ?) de réponse

Réponses regroupés dans l'espace vectoriel (corpus – 2nd ordre)







Bilan et perspectives

- Des mesures qui permettent de
  - "scorer" la proximité entre un stimulus et une réponse
  - qualifier la relation et les réponses récoltées
  - identifier des groupes de réponses ... des groupes de répondants ?
- De nouvelles données récoltées (Evolex v2)
  - Corrélation entre temps de réponse et "profils" de réponse
- Des données à récolter auprès de personnes atteintes d'un trouble du langage
  - Corrélation entre "profils" de réponse et pathologies développementale, dégénérative et lésionnelle