EVOLEX : approches psycholinguistique et computationnelle de l'accès au lexique et de la proximité sémantique entre paires de mots

Xavier de Boissezon, Lola Danet, Cécile Fabre, Jérôme Farinas, Bruno Gaume, Nabil Hathout, Lydia-Mai Ho-Dac, Mélanie Jucla, Patrice Péran, Bénédicte Pierrejean, et al.

EVOLEX : APPROCHES PSYCHOLINGUISTIQUE ET COMPUTATIONNELLE DE L'ACCÈS AU LEXIQUE ET DE LA PROXIMITÉ SÉMANTIQUE ENTRE PAIRES DE MOTS

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

[1]ToNIC, Inserm, UT3 [2]CLLE-ERSS, CNRS, UT2J [3]IRIT, CNRS, UT3 [4]Octogone-Lordat, UT2J **Contact:** xavier.deboissezon@inserm.fr, melanie.jucla@univ-tlse2.fr, ludovic.tanguy@univ-tlse2.fr, evolex@irit.fr

Objectifs

Développer une **méthode assistée par ordinateur** pour évaluer auprès de populations avec ou sans troubles du langage la façon dont

- nous accédons aux informations présentes dans notre lexique mental
- nous associons ces informations les unes avec les autres

Pour réaliser cet objectif, nous allions des compétences en psycholinguistique et neuropsychologie d'une part et de traitement automatique des langues naturelles d'autre part

Organisation du projet EVOLEX

1. Développement d'outils pour la reconnaissance automatique de la parole et la correction manuelle des données récoltées (productions orales et temps de réponse) IRIT, ToNIC
2. Définition des épreuves, du matériel linguistique utilisé, collecte et prétraitement des données Octogone-Lordat, ToNIC
3. Annotation et analyses qualitatives des réponses CLLE-ERSS, Octogone-Lordat
4. Application des modèles de Sémantique Distributionnelle pour évaluer les données CLLE-ERSS
5. Mise en commun des analyses pour la caractérisation de profils de locuteurs et des réseaux sémantiques CLLE-ERSS, Octogone-Lordat, ToNIC

Protocole EVOLEX

3 épreuves d'accès lexical dont une épreuve de **G**énération **V**erbale (GV) :
1 nom entendu ⇒ le premier mot qui vient à l'esprit (ex : *abricot* → *confiture*)

- Recueil des données (logiciel Samoplay)

- Passations auprès de locuteurs avec ou sans troubles du langage de 18 à 90 ans

Ce projet a bénéficié du soutien de l'Institut des Handicaps du CHU de Toulouse et de la Fédération Hospitalo-Universitaire des Handicaps Cognitifs, Psychiques et Sensoriels

Références

[1] B. Gaume, K. Duvignau, E. Navarro, Y. Desalle, H. Cheung, S.-K. Hsieh, P. Magistry, and L. Prevot.
Skillex : a graph-based lexical score for measuring the semantic efficiency of used verbs by human subjects describing actions.
Revue TAL, 55(3), 2016.

[2] B. Gaume, L. Tanguy, C. Fabre, L.-M. Ho-Dac, B. Pierrejean, N. Hathout, J. Farinas, J. Pinquier, L. Danet, P. Péran, X. De Boissezon, and M. Jucla.
Automatic analysis of word association data from the evolex psycholinguistic tasks using computational lexical semantic similarity measures.
In 13th International Workshop on Natural Language Processing and Cognitive Science (NLPCS), Krakow, Poland, 2018.

[3] P. Péran, J.-F. Démonet, C. Pernet, and D. Cardebat.
Verb and noun generation tasks in huntington's disease.
Movement disorders : official journal of the Movement Disorder Society, 19(5) :565–571, 2004.

[4] L. Tanguy, F. Sajous, and N. Hathout.
Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques.
Traitement Automatique des Langues, 56(2), 2015.

Participation de quatre laboratoires toulousains

CLLE-ERSS est une équipe de linguistique comportant un axe de recherche en traitement automatique des langues

SAMoVA de l'IRIT pour « **S**tructuration, **A**nalyse et **Mo**délisation de documents **V**idéo et **A**udio »

Octogone-Lordat, laboratoire de neuro-psycholinguistique, mène des recherches sur l'étude des (dys)fonctionnements langagiers et de la cognition

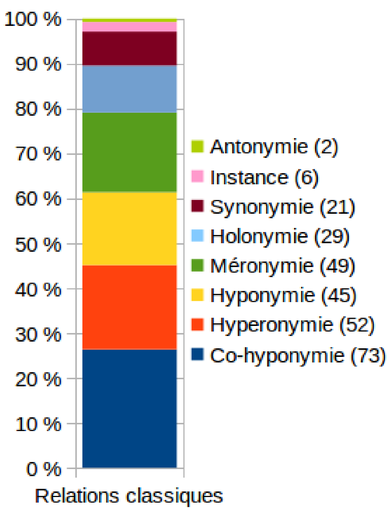
ToNIC pour « Toulouse Neuro Imaging Center », a pour objectif principal l'étude du cerveau humain et des principales pathologies qui l'affectent

Données (GV, version 1, 30 sujets)

- 1800 paires stimulus-réponse récoltées
- 1544 paires exploitables (559 différentes)

Double-annotation des relations

⇒ des **relations "classiques"** (50%) :



⇒ des **relations "autres"** : association d'idées (36%), syntagmes (9%), pas de relation (5%), lien phonologique (1%)

Méthode

4 mesures pour évaluer la similarité / proximité sémantique entre mots

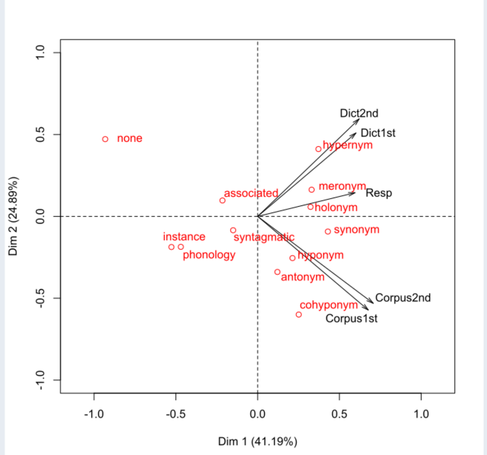
2 similarités issue de corpus

- ***Corpus1st*** : Plus deux mots apparaissent fréquemment ensemble en corpus, plus ces deux mots sont similaires (= collocations)
- ***Corpus2nd*** : : Plus deux mots partagent les mêmes collocations, plus ces deux mots sont similaires (= substituabilité)
- corpus = FrWac (pages Web *.fr*, 2 milliards de mots)

2 similarités issues de dictionnaires

- ***Dict1st*** : Si un mot apparaît dans la définition d'un autre mot, ces deux mots sont similaires
- ***Dict2nd*** : Plus les définitions de deux mots partagent de mots, plus ces deux mots sont similaires
- dict. = TLF – Trésor de la Langue Française

Résultats principaux



Analyse en Composantes Principales

→ Des mesures positivement corrélées aux **relations classiques**

→ Des mesures négativement corrélées dans les cas de
- sens éloigné (sans relation identifiée, phonologie : *chou* – *caillou*)
- d'instance (*magicien* – *Merlin*)

→ Des aspects différents captés par les différentes méthodes
- **dictionnaire** – hyperonymie (*entrecôte* – *viande*)
- **corpus** – cohyponymie (*balançoire* – *toboggan*)

→ Association et Syntagmatique non corrélées et caractérisées

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

Identification de profils de réponses