Table ronde RISE 2018

Enjeux et perspectives de la recherche d'information sémantique - Ressources lexicales

Didier Schwab

Qu'est-ce que le « sens » ?

- Ce qui va me permettre de résoudre un certain nombre d'ambiguïtés dans un énoncé

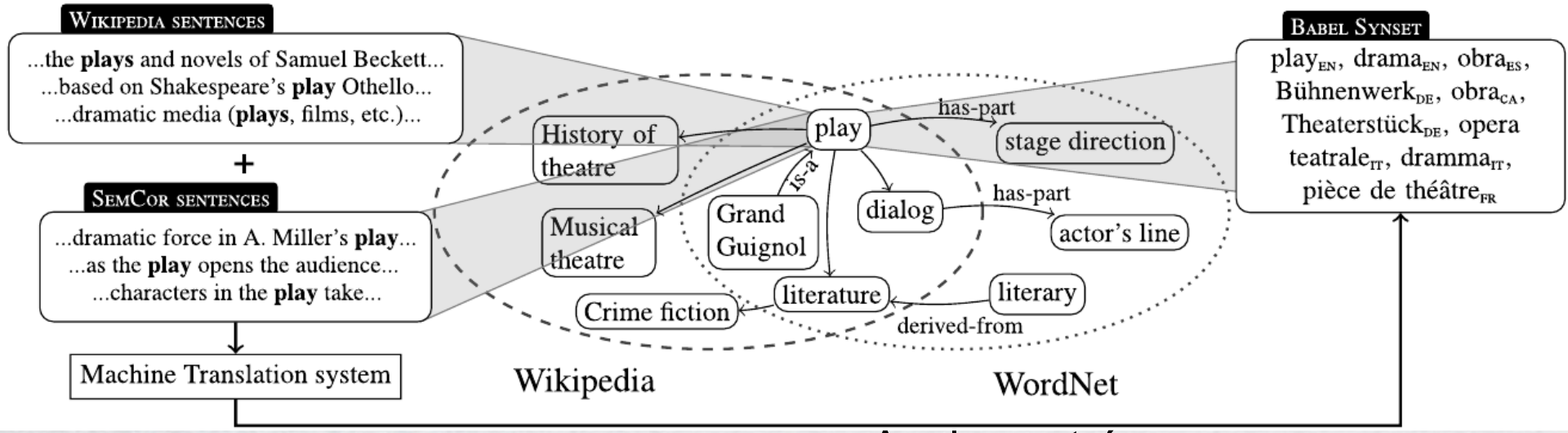
  La souris est montée sur la table. Elle a mangé le câble de l'ordinateur.

- Nécessite des connaissances implicites/explicites
  - Une *souris/animal* peut monter ; la *souris/terminal* moins facilement
  - Câble/électronique plus proche d'ordinateur que Câble/téléphérique
- Connaissances explicites
  - Bases lexicales : wordnets, JeuxDeMots, BabelNet (wordnets+Wikipedia+…), DBNary (extraction Web Sémantique wiktionnaire pour 16 langues)
- Connaissances implicites
  - Thésaurus distributionnels, vecteurs,

Princeton WordNet de l'anglais

- Travail conséquent réalisé manuellement 1985 → 2006
  [Miller, 1995] [Fellbaum, 1998]
- Dictionnaire + réseau lexical entre sens
- Très riche écosystème
  - corpus annotés en sens
  - Lien avec Ontologies (Sumo)
  - Liens avec d'autres ressources lexicales y compris dans d'autres langues
- Wordnets (*Open Multilingual Wordnet*) :
  - 34 langues
  - Anglo-centré : Perte d'informations → pis-aller

BabelNet [Navigli & Ponzetto, 2012]

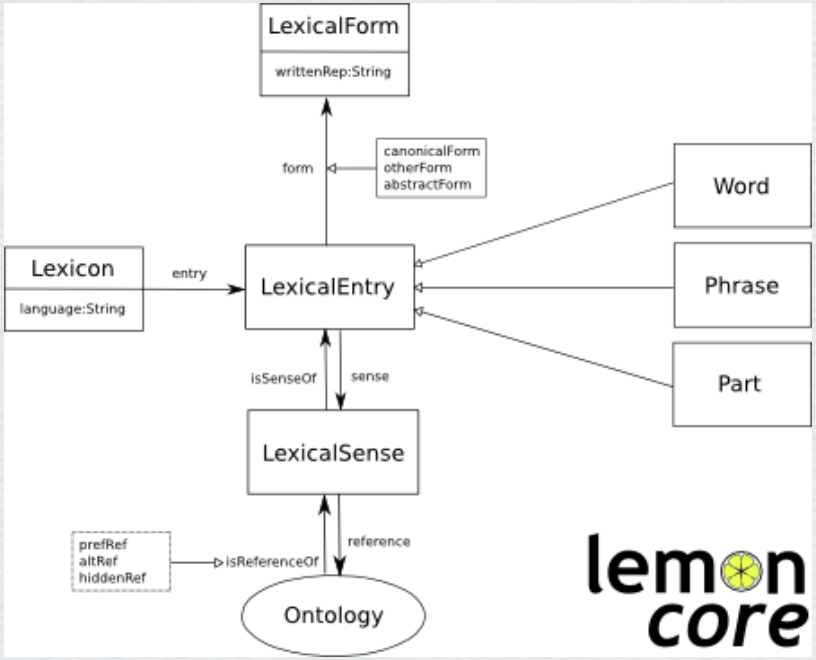


- 2011-...
- 284 langues
- gain qualitatif de la désambiguïsation lexicale pour l'anglais vers 2013-2014

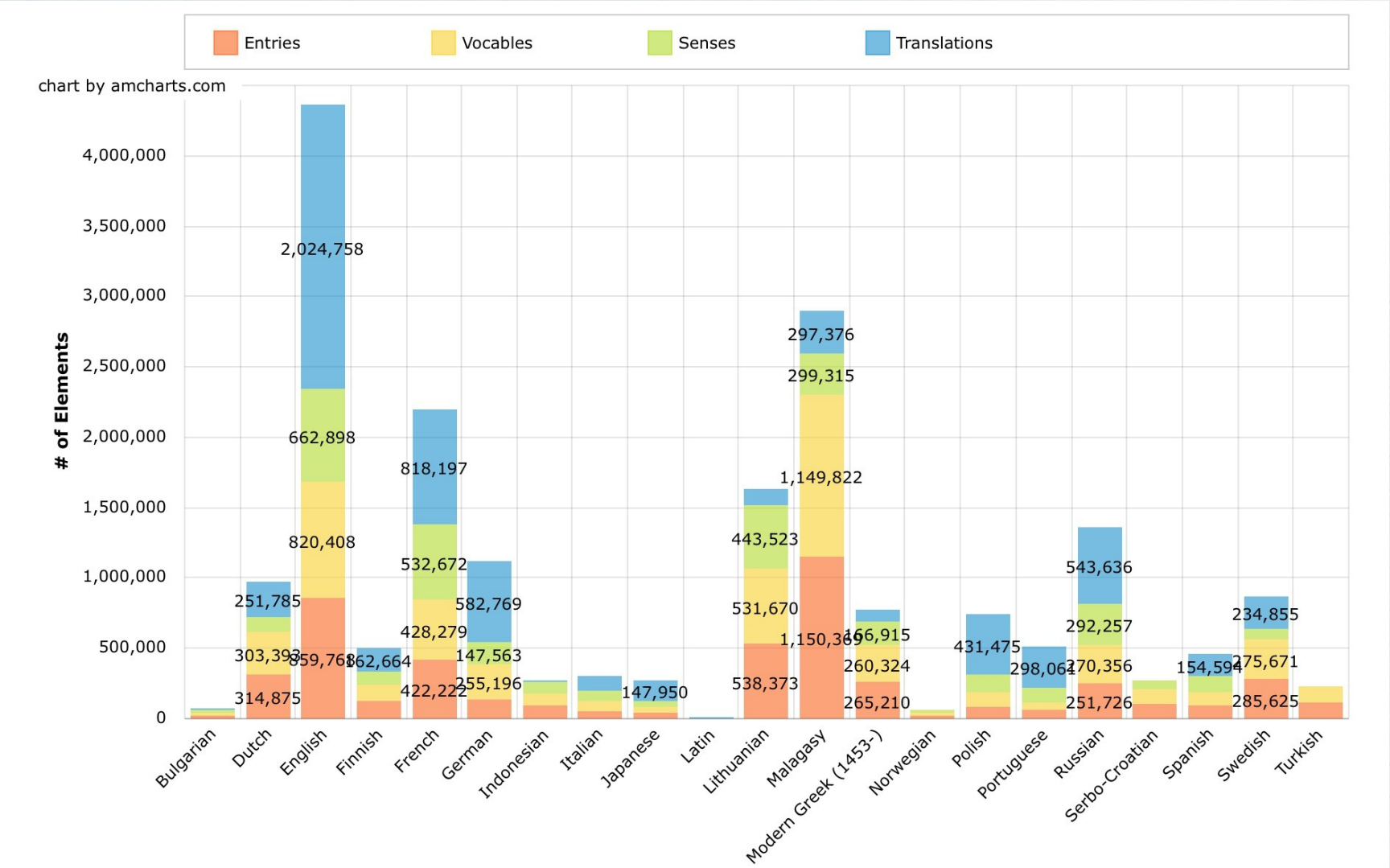
- Anglo-centré
- Principalement les noms
- Alignement automatisé
- Qualité des autres langues que l'anglais ?
- Licence de plus en plus restrictive

DBNary [Sérasset, 2014]

- Gilles Sérasset 2011-...
- Extraction du Wiktionaire pour 21 langues
- RDF – modèle lemon (Lexicon Model for Ontologies)
- Entrées, sens, traductions
- Non anglo-centré (axis)
- Mais nécessite un convertisseur par langue



DBNary


Entries
Vocables
Senses
Translations
chart by amcharts.com
# of Elements
4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
Bulgarian
Dutch
English
Finnish
French
German
Indonesian
Italian
Japanese
Latin
Lithuanian
Malagasy
Modern Greek (1453-)
Norwegian
Polish
Portuguese
Russian
Serbo-Croatian
Spanish
Swedish
Turkish
314,875
303,393
251,785
859,768
662,898
820,408
2,024,758
662,664
422,222
428,279
532,672
818,197
255,196
147,563
582,769
147,950
538,373
531,670
443,523
1,150,369
299,315
1,149,822
297,376
265,210
260,324
166,915
431,475
298,064
251,726
270,356
292,257
543,636
154,594
285,625
275,671
234,855

Représentations distribuées (vectorielles)

- Componentielles
  - Composantes des vecteurs représentent des idées, concepts…
    - Directement : Vecteurs sémantiques 1992 → ~ 2005 [Chauché, 2003]
    - Indirectement : Vecteurs conceptuels 1997 → ~ 2012 [Lafourcade, 2001] [Schwab, 2005]
  - Représentation au niveau du sens (dictionnaires), au niveau du mot (fusion des sens : sommes pondérées)
- Distributionelles
  - Composantes des vecteurs représentent les voisins dans un corpus
    - Directement : Salton
    - Indirectement : LSA, Words Embeddings (Word2Vect, Glove)
  - Représentation au niveau du mot, recherche assez foisonnante pour la représentation au niveau du sens et de la phrase, texte
  - Monolingue - Multilingue

Représentations au niveau de la phrase, du syntagme, du texte

- Beaucoup se contentent de la somme des mots du vecteur
- Méthode de [Ferrero et al., 2017a] (1er SemEval 2017 [Ferrero et al., 2017b]) inspirée de
  - [Lafourcade, Schwab] pondération en fonction des parties du discours
  - [Brychcin & Svoboda, 2016] pondération en fonction de la fréquence inverse en documents
- Corrélation avec jugement humain [Billah Nagoudi et al., 2017]

Corpus	MSRpar	MSRvid	SMTeuroparl	STS 2017	Global
Pondération unitaire	0,6745	0,7233	0,6233	0,5957	0,6653
Pondération avec IDF	0,7432	0,7820	0,7110	0,7309	0,7467
Pondération avec POS	0,7446	0,7951	0,7317	0,7425	0,7562
Pondération mixte	**0,7523**	**0,8276**	**0,7460**	**0,7646**	**0,7745**

Représentations des sens

- Induction de sens (WSI)
  - Grands corpus de textes -> inférer les sens possibles pour chacun des mots qui le composent
- Sens définis *a priori*
  - Approche dictionnairique
    - [Ferrero et al., 2017a] inspiré de [Ferrero, 2017] et [Schwab, Lafourcade] : Vecteur de chaque sens calculé grâce aux définitions dans WordNet
  - Approche corpus
    - Corpus Annotés en sens issus de WordNet
    - Apprentissage à partir du voisinage
    - Word2Vec [Iacobacci et al., 2015]

Format d'unification des corpus annotés en sens : UFSAC

- Plus d'une dizaine de corpus annotés en sens WordNet
  - Versions différentes de WordNet
  - Formats différents
  - Informations différentes
- UFSAC
  - [Vial et al., 2018a]
  - Unified Format for Sense Annotated Corpora
  - WordNet 3.0
  - Informations unifiées

```
<corpus id="short_example">
  <document id="d001" >
    <paragraph>
      <sentence>
        <word surface_form="A" pos="DT" />
        <word surface_form="precise"
           wn30_key="precise%3:00:00::" />
        <word surface_form="example"
              pos="NN" lemma="example" />
        <word surface_form="." />
      </sentence>
    </paragraph>
  </document>
</corpus>
```

Format d'unification des corpus annotés en sens : UFSAC

Corpus	Sentences	Words		Annotated parts of speech			
		Total	Annotated	Nouns	Verbs	Adj.	Adv.
SemCor	37176	778587	229517	87581	89037	33751	19148
DSO	178119	5317184	176915	105925	70990	0	0
WordNet GlossTag	117659	1634691	496776	232319	62211	84233	19445
MASC	34217	596333	114950	49263	40325	25016	0
OMSTI	820557	35843024	920794	476944	253644	190206	0
Ontonotes	21938	435340	52263	9220	43042	0	0
SemEval 2007 task 07	245	5637	2261	1108	591	356	206
SemEval 2007 task 17	120	3395	455	159	296	0	0
SemEval 2013 task 12	306	8142	1644	1644	0	0	0
SemEval 2015 task 13	138	2638	1053	554	251	166	82
Senseval 2	238	5589	2301	1061	541	422	277
Senseval 3 task 1	300	5511	1957	886	723	336	12

UFSAC dans d'autres langues

- La plupart des langues ont peu de données annotées en sens librement disponibles
- Français 5000 mots ? Arabe 14000 Ontonote [Hadj-Salah, 2018a]
- Traduction des corpus UFSAC et portage des annotations [Hadj-Salah, 2018b]
- UFSAC-ara, UFSAC-fra,…
- Désambiguïsation lexicale, Embeddings de sens translingues,...

UFSAC pour la désambiguïsation lexicale

Système	SE2	SE3	SE07 (07)	SE07 (17)	SE13 (12)	SE15 (13)
Notre système	**73.75%**	70.31%	**83.59%**	60.22%	68.98%	*73.98%
(Yuan et al., 2016) (LSTM)	73.6%	69.2%	82.8%	**64.2%**	67.0%	72.1%
(Yuan et al., 2016) (LSTM + LP)	**73.8%**	**71.8%**	**83.6%**	63.5%	**69.5%**	**72.6%**
(Raganato et al., 2017) (BLSTM)	71.4%	68.8%	-	*61.8%	65.6%	69.2%
(Raganato et al., 2017) (BLSTM + att. + LEX + POS)	72.0%	69.1%	83.1%	*64.8%	66.9%	71.5%
Sens le plus fréquent	65.6%	66.0%	78.89%	54.5%	63.8%	67.1%
(Iacobacci et al., 2016)	68.3%	68.2%	-	59.1%	-	-

- Système de DL neuronale anglais entraîné sur les 6 corpus UFSAC anglais (SemCor, DSO, WNGT, MASC, OMSTI et OntoNotes anglais)

[Vial et al., 2018b]
[Hadj Salah et al., 2018a et b]

	Précision	Rappel	Score F1
Dev (SE15 traduit automatiquement de l'anglais)			
Notre système	70.06%	69.55%	69.81%
Notre système + repli premier sens	70.28%	70.28%	70.28%
Baseline aléatoire	36.92%	36.92%	36.92%
Baseline du sens le plus fréquent	67.25%	67.25%	67.25%
Test (OntoNotes arabe)			
Notre système	**71.52%**	**71.32%**	**71.42%**
Notre système + repli premier sens	**71.52%**	**71.32%**	**71.42%**
Baseline aléatoire	40.26%	40.04%	40.15%
Baseline du sens le plus fréquent	58.11%	57.95%	58.03 %

- Système de DL neuronale arabe entraîné sur les 12 corpus UFSAC arabes

Futur ?

- Augmenter la quantité de corpus annotés
  - Pas forcément naïvement
  - Comprendre pourquoi certaines mauvaises annotations
  - Savoir choisir les termes à annoter pour sélectionner les exemples adéquats
  - Mesures de confiance

→ améliorer la désambiguïsation

→ créer des vecteurs multilingues

Bibliographie

[Miller, 1995] George A. Miller. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[Fellbaum, 1998] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[Navigli & Ponzetto, 2012] R. Navigli and S. Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250

[Sérasset, 2014] Sérasset Gilles (2014). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. to appear in Semantic Web Journal (special issue on Multilingual Linked Open Data).

Making people play for Lexical Acquisition with the JeuxDeMots prototype M Lafourcade SNLP'07: 7th international symposium on natural language processing

[Lafourcade, 2001] Lexical sorting and lexical transfert by conceptual vectors. In proc. of the First International Workshop on MultiMedia Annotation (MMA'2001) Tokyo, January 2001, 6p.

[Chauché, 2003] Jacques Chauché , Violaine Prince Simon Jaillet, Maguelonne Teisseire Classification automatique de textes à partir de leur analyse syntaxico-sémantique, TALN 2003, Batz-sur-Mer, 11–14 juin 2003

[Schwab, 2005] Didier Schwab. Approche hybride - lexicale et thématique - pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte. Interface homme-machine [cs.HC]. Université Montpellier II - Sciences et Techniques du Languedoc, 2005.

[Ferrero et al., 2017a] Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, Didier Schwab. Using Word Embedding for Cross-Language Plagiarism Detection. EACL 2017, Apr 2017, Valence, Spain. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2,, 2, pp.415 - 421.

[Ferrero et al., 2017b] Jérémy Ferrero, Laurent Besacier, Didier Schwab, Frédéric Agnès. CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity. Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017),, Aug 2017, Vancouver, Canada. Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017),, 2017.

Bibliographie

[Brychcin & Svoboda, 2016] Tomas Brychcin and Lukas Svoboda. 2016. UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic in formation. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016). San Diego, CA, USA, pages 588– 594.

- [Billah Nagoudi et al., 2017] El Moatez Billah Nagoudi, Jérémy Ferrero, Didier Schwab. Amélioration de la similarité sémantique vectorielle par méthodes non-supervisées. 24e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017), Jun 2017, Orléans, France.

[Vial et al., 2017] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Sense Embeddings in Knowledge-Based Word Sense Disambiguation. 12th International Conference on Computational Semantics, Sep 2017, Montpellier, France.

[Iacobacci et al., 2015] Ignacio Iacobacci, Mohammad Taher Pilehvar and Roberto Navigli SENSEMBED: Learning Sense Embeddings for Word and Relational Similarity Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 95–105, Beijing, China, July 26-31, 2015.

[Vial et al., 2018a] Loïc Vial, Benjamin Lecouteux, Didier Schwab. UFSAC: Unification of Sense Annotated Corpora and Tools. Language Resources and Evaluation Conference (LREC), May 2018, Miyazaki, Japan.

[Vial et al., 2018b] Loïc Vial, Benjamin Lecouteux, Didier Schwab. Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale. 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France.

[Hadj Salah, 2018a] Marwa Hadj Salah, Loïc Vial, Hervé Blanchon, Mounir Zrigui, Benjamin Lecouteux, et al.. La désambiguïsation lexicale d'une langue moins bien dotée, l'exemple de l'arabe. 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France.

[Hadj Salah, 2018b] Marwa Hadj Salah, Hervé Blanchon, Mounir Zrigui, Didier Schwab. Un corpus en arabe annoté manuellement avec des sens WordNet. 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France.