Fr-Semcor : annotation sémantique à gros grain

Lucie Barque

Université Paris 13 & LLF, Université Paris 7 & CNRS

Avec : M. Candito, B. Crabbé, P. Haas, R. Huyghe, H. Martínez Alonzo, D. Tribout

Séminaire ERSS

21 Décembre 2017, Toulouse

Introduction

- Fr-Semcor : projet initié en juin 2017 dans le cadre du Labex EFL
- Annotation sémantique manuelle d'un corpus du français
  - Pour l'instant limitée aux noms communs
  - À l'aide de classes sémantiques générales (gros grain)
- Objectif : contribuer au développement d'un outil d'analyse sémantique du français en fournissant des données pour l'apprentissage semi-supervisé
  - Désambiguïsation nominale
  - Désambiguïsation verbale

Plan

Positionnement

Annotation à gros grain : deux configurations possibles

1. On dispose d'un lexique dans lequel les sens des mots sont listés
   - Réduction de la polysémie (Navigli 2006, Palmer *et al.* 2007)

     Ex. BASS : 8 synsets, 6 *Unique Beginners* distincts
     → Choix d'un de ces 6 "supersens" pour une occurrence de BASS

2. On ne dispose pas d'un tel lexique
   - Inventaire de classes sémantiques générales (Ciaramita et Johnson, 2003, Schneider *et al.* 2012, Pederson *et al.* 2016)

     Ex. 25 *Unique Beginners* (Animal, Person, Body, Act, Attribute, etc.)
     → Choix d'une de ces 25 "super-classes" pour une occurrence de BASS

☞ Granularité sémantique potentiellement différente dans les deux cas

☞ Tâche d'annotation distincte

Annotation à gros grain : Justifications

- Difficile d'annoter avec un grain fin
  - Accord inter-annotateur avec WordNet : $\sim$ 70% (Snyder and Palmer 2004)
  - Accord inter-annotateur avec un gros grain : $\sim$ 94% (Navigli *et al.* 2007)
- Le gros grain requiert un nombre moins important d'annotations pour l'apprentissage
  - P. ex. plus d'occurrences de N annotées Artefact que d'occurrences de N annotées Meuble
- Le gros grain est suffisant pour plusieurs tâches
  - Restriction de sélection des arguments des verbes

Positionnement

Corpus existants : annotés manuellement

- Anglais
  - SemCor 3.0 (Landes *et al.* 1998)
  - Ontonotes 5.0 (Weischedel *et al.* 2013)
  - Environ 250 000 occurrences annotées (N, V, Adj)
  - Inventaire de WordNet (grain + ou - fin)

- Autres langues
  - Sur le modèle du Princeton SemCor
    - Langues : bulgare, basque, espagnol, catalan, japonais, hollandais, allemand, italien, hongrois, polonais, etc.
    - Voir (Petrolino & Bond 2014)
  - Inventaire à gros grain développé à partir des WordNet *Unique Beginners*
    - Langues : arabe (Schneider *et al.* 2012), danois (Pederson *et al.* 2016)
    - Environ 35 000 occurrences annotées
  - ☞ Pas de ressources comparables pour le français

Rappel : découpage de la polysémie dans WordNet

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: bass [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: <lexical filename > (gloss) "an example sentence"

Noun

- <noun.attribute>S: (n) **bass** (the lowest part of the musical range)
- <noun.communication>S: (n) **bass**, bass part (the lowest part in polyphonic music)
- <noun.person>S: (n) **bass**, basso (an adult male singer with the lowest voice)
- <noun.food>S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- <noun.food>S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- <noun.communication>S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- <noun.artifact>S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- <noun.animal>S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Positionnement

Corpus existants : annotés automatiquement

- Multilingue : EuroSense (Delli Bovi *et al.* 2017)
  - Annotation sémantique automatique d'Europarl
  - 123 millions d'occurrences annotées à l'aide de 155 000 sens (concept + EN) issus de BabelNet
  - Français $\rightarrow$ env. 13 millions d'occurrences annotées
    - Précision : 71,8 %
    - Couverture : 63,5 %

- Bien que bruités, ces corpus permettent d'améliorer la WSD

Plan

Définition des classes sémantiques

Les *Unique Beginners*

- Etiquettes issues du projet WordNet (Miller *et al.* 1990, Fellbaum 1998)
  - Act, Animal, Artifact, Attribute, Body, Cognition, Communication, Event, Feeling, Food, Group, Location, Motive, Object, Person, Phenomenon, Plant, Possession, Process, Quantity, Relation, Shape, State, Substance, Time, Tops
  - ☞ Pas un jeu d'étiquettes construit pour l'annotation mais pour le travail lexicographique

Définition des classes sémantiques

Les *Unique Beginners*

- Etiquettes issues du projet WordNet (Miller *et al.* 1990, Fellbaum 1998)
  - Act, Animal, Artifact, Attribute, Body, Cognition, Communication, Event, Feeling, Food, Group, Location, Motive, Object, Person, Phenomenon, Plant, Possession, Process, Quantity, Relation, Shape, State, Substance, Time, Tops
  - ☞ Pas un jeu d'étiquettes construit pour l'annotation mais pour le travail lexicographique

- Classes hétérogènes
  - En termes ontologique (Act, Person vs Cognition, Communication)
  - En termes de taille (Act, Person vs Shape, Motive)

Définition des classes sémantiques

Les *Unique Beginners*

- Etiquettes issues du projet WordNet (Miller *et al.* 1990, Fellbaum 1998)
  - Act, Animal, Artifact, Attribute, Body, Cognition, Communication, Event, Feeling, Food, Group, Location, Motive, Object, Person, Phenomenon, Plant, Possession, Process, Quantity, Relation, Shape, State, Substance, Time, Tops
  - ☞ Pas un jeu d'étiquettes construit pour l'annotation mais pour le travail lexicographique

- Classes hétérogènes
  - En termes ontologique (Act, Person vs Cognition, Communication)
  - En termes de taille (Act, Person vs Shape, Motive)

- Un principe : classes complémentaires
  - Ex. Bien que des N étiquetés Food correspondent également à des substances, ils ne sont pas étiquetés Substance
  - = Substance : tous les N qui dénotent des substances excepté les substances de type aliment

Définition des classes sémantiques

Les *Unique Beginners*

- Définitions sommaires

noun.group	nouns denoting groupings of people or objects
noun.location	nouns denoting spatial position
noun.motive	nouns denoting goals
noun.object	nouns denoting natural objects (not man-made)
noun.person	nouns denoting people
noun.phenomenon	nouns denoting natural phenomena
noun.plant	nouns denoting plants
noun.possession	nouns denoting possession and transfer of possession
noun.process	nouns denoting natural processes
noun.quantity	nouns denoting quantities and units of measure
noun.relation	nouns denoting relations between people or things or ideas
noun.shape	nouns denoting two and three dimensional shapes
noun.state	nouns denoting stable states of affairs
noun.substance	nouns denoting substances
noun.time	nouns denoting time and temporal relations

Définition des classes sémantiques

Adaptation des *Unique Beginners* pour Fr-SemCor

- Suppression d'étiquettes
  - Ex. abandon de l'étiquette Location
    - Lieu naturel (ex. *plage, campagne, terrain*) → Object
    - Lieu construit (ex. *salon, ville*) → Artifact
    - N de localisation interne (ex. *coin, bord*) → Part
- Ajout d'étiquettes (Institution et Part)
- Redéfinition d'étiquettes
  - Ex. l'étiquette Cognition
    - Cognition WordNet : → Inclut les actes cognitifs
    - Cognition dans FR-SemCor : → N'inclut pas les actes cognitifs (dans Act)

Classes sémantiques

Adaptation des *Unique Beginners* pour Fr-SemCor

- Rédaction d'un guide d'annotation
  - Définition des classes
    - Ex. Attribute : propriété constitutive d'une entité (individu, objet, situation) ou propriété en soi
  - Tests indicatifs
    - Ex. Attribute : *X est d'un grand N, X a fait preuve de N/d'un N exp*
  - Exemples d'occurrences annotées
    - *De **mémoire**$_{Attribute}$ de météorologues, seuls l'ont surpassé des typhons extrêmes dans le Pacifique.*
    - *Ils s'éloignent lentement en direction de la passe du Lido, pour retrouver l'**immensité**$_{Attribute}$ de la mer Adriatique.*
    - *"..." se souvient la jeune femme, **silhouette**$_{Attribute}$ fine et yeux sombres.*
    - *Ils s'agenouillent pendant l'hymne américain pour dénoncer le **racisme**$_{Attribute}$.*

Classes sémantiques

Adaptation des *Unique Beginners* pour Fr-SemCor

- Explicitation des répartitions entre classes. Exemples :
  - Les bactéries et les virus sont étiquetés Substance (et non Animal)
  - Les aliments sont étiquetés Food (et non Substance)
  - Les fluides corporels sont étiquetés Substance (et non Body)
  - Les fruits, légumes et céréales sont étiquetés Plant (et non Food)
  - Les unités de mesure de temps sont étiquetées Time (et non Quantity)
  - Les monnaies sont étiquetées Possession (et non Quantity)

Classes sémantiques

Classes du projet FR-SemCor

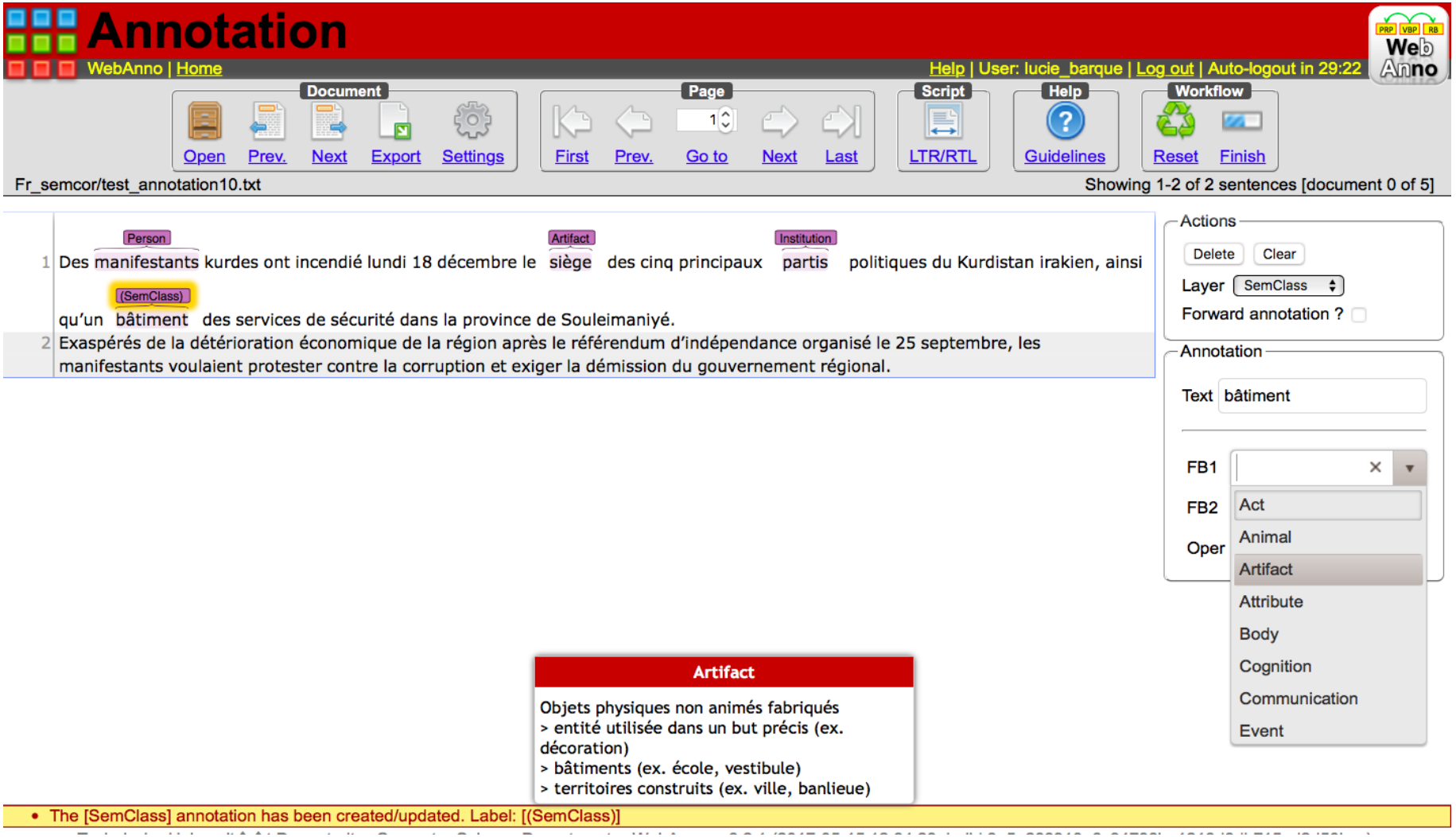
1. Classes complémentaires
   - Entités concrète : Animal, Person, Plant, Artifact, Body, Food, Object, Substance
   - Situations dynamiques : Act, Event
   - Situations statives ou propriétés : Attribute, State, Feeling
   - Autres : Quantity, Institution, Cognition, Relation, Possession, Phenomenon, Shape, Time
2. Classes relationnelles
   - Groupes ou parties (de qqch) : Group, Part
3. Classe pour les N sous-spécifiés : Tops

Corpus

Le corpus Sequoia

- Corpus du français
  - $\sim$ 3100 phrases (67 000 tokens)
  - Issues de Europarl, Est Républicain, Wikipedia-fr et EMEA
- Couches d'annotations manuelles
  - Syntaxe de surface (Candito et Seddah, 2012)
  - Syntaxe profonde (Candito *et al.*, 2014)
  - Frames et Frame Elements (Djemaa *et al.*, 2016)
  - Expressions polylexicales (Parseme-fr)
- Données nominales
  - 2786 noms communs, $\sim$ 13 000 tokens
  - 634 noms > 5 occurrences

L'outil WebAnno (Eckart de Castilho *et al.*, 2016)



Plan

Annotation : phase test

Accord (Kappa)

- 300 occurrences de N provenant de 9 textes extraits du journal *Le Monde*

	Anno 1	Anno 2	Anno 3	Anno 4
Anno 1	—-	0,8	0,74	0,77
Anno 2	0,8	—-	0,75	0,78
Anno 3	0,74	0,75	—-	0,72
Anno 4	0,77	0,78	0,72	—-

Annotation : phase test

Problèmes récurrents

- Définition des classes
- Repérage et annotation des expressions polylexicales
- Interprétation contextuelle
  - L'ambiguïté en contexte
  - L'implicite

Annotation : phase test

Définition des classes : problèmes de fontière

- Ex. Act et Event : est-ce que le N décrit une action qui a un "effectuateur", i.e quelque chose ou quelqu'un qui déploie l'énergie qui permet de réaliser l'action ?

  - *On remarque immédiatement la **formation** d'un empois d'amidon.*
  - *Absent lors de l'**audience** de vendredi, il a été reconnu coupable du meurtre de sa petite amie.*
  - *La gestion du cheptel humain s'est matérialisée dans diverses technologies politiques qui ont savamment entravé ou au contraire, encouragé sa **reproduction** selon les cours, les flux et les circuits de la traite.*

☞ Opérateur / en cas d'ambiguïté

Annotation : phase test

Définition des classes : problèmes de fontière

- Ex. Act et Event : est-ce que le N décrit une action qui a un "effectuateur", i.e quelque chose ou quelqu'un qui déploie l'énergie qui permet de réaliser l'action ?

  - *On remarque immédiatement la **formation**$_{Event}$ d'un empois d'amidon.*
  - *Absent lors de l'**audience**$_{Act}$ de vendredi, il a été reconnu coupable du meurtre de sa petite amie.*
  - *La gestion du cheptel humain s'est matérialisée dans diverses technologies politiques qui ont savamment entravé ou au contraire, encouragé sa **reproduction**$_{Act\ /\ Event}$ selon les cours, les flux et les circuits de la traite.*

☞ Opérateur / en cas d'ambiguïté

Annotation : phase test

Définition des classes : problèmes de couverture

- Noms difficiles à classer
  - *La France a constamment activé dans son propre **intérêt** ce mythe de la surpopulation des territoires colonisés.*
    - Cognition / Relation / Attribute / Possession
  - *Il a toujours plaidé la **méprise**, affirmant qu'il était persuadé qu'un cambrioleur s'était introduit dans sa propriété de Pretoria et qu'il avait tiré sous le coup de la panique.*
    - Act / State / Event / Act / Feeling

Annotation : phase test

Les expressions polylexicales (EP)

- Repérage : les EP sont annotées dans le Sequoia (projet Parseme-fr)
- Annotation
  - Pas d'annotation des N dans des EP autres que nominales
    - *sous le coup de*, *en effet*, *à l'origine de*
  - Annotation des EP nominales dans leur entier, quel que soit leur degré de compositionnalité sémantique
    - [*cul de poule*]$_{Artifact}$
    - [*garçon de café*]$_{Person}$

Annotation : phase test

Interprétation contextuelle

- L'ambiguïté en contexte : deux opérateurs

  1. L'opérateur de conjonction **+** : dans le contexte, l'occurrence du nom renvoie explicitement à A et à B (coprédication)

  - *Selon une **étude**$_{Act+Cognition}$ menée par l'UNAF, une banque prélève chaque année en moyenne 34 euros de frais pour incident de paiement.*
  - *La répétition du passage de ces monstres des mers entraîne des **dommages**$_{Artifact+Event}$ irréparables sur les fondations des palais et des églises qu'ils frôlent.*

  2. L'opérateur de disjonction **/** : le contexte ne permet pas de décider si le nom renvoie à A ou à B ou à A et B

  - *Son **mariage**$_{Event\,/\,State}$ repose sur un mensonge.*
  - *Elles ne bénéficient pas de **traitements**$_{Act\,/\,Artifact}$ efficaces et peuvent être à l'origine de nouvelles contaminations sans le savoir*

☞ Pas de décision *a priori* sur les N à facettes (Cruse 1995)

Annotation : phase test

Interprétation contextuelle

- Le cas des N d'ensemble

  N généraux dénotant un groupe
  - *Le module distant sans fil comprend une antenne et un* ***ensemble***$_{Group}$ *de commutateurs*$_{Artifact}$*.*
  - *L'infirmière n'est pas dans la salle, avec le* ***troupeau***$_{Group}$ *des débatteurs*$_{Person}$ *: elle les domine.*

  N qui spécifient lexicalement les éléments regroupés
  - ☞ Opérateur de distribution X
  - *Une banque prélève chaque année en moyenne 34 euros de frais pour incident de paiement sur l'ensemble de sa* ***clientèle***$_{Group X Person}$*.*

  N généraux avec éléments regroupés inférables du contexte
  - ☞ Opérateur de distribution X
  - *Le* ***groupe***$_{Group X Person}$ *marchait en silence dans la montagne.*

Conclusion

Projet encore au stade de la stabilisation du guide d'annotation

La suite

- Annoter les 13 000 N du Sequoia
  - Annotation par N
    - Pré-annotation automatique pour les N monosémiques
    - Ex. Les noms PATIENT (233 occ), AN (120 occ), MILLIGRAMME (80 occ)
  - Quatre annotateurs experts + deux stagiaires
- Apprentissage semi-supervisé à partir des données

Références 1

- Candito M. and Seddah D. (2012) Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *Actes de TALN'2012*, Grenoble, France
- Candito M., Perrier G., Guillaume B., Ribeyre C., Fort K., Seddah D. and de la Clergerie E. (2014) Deep Syntax Annotation of the Sequoia French Treebank. *Proceedings of LREC 2014*, Reykjavic, Iceland.
- Ciaramita M. and M. Johnson (2003) Supersense Tagging of Unknown Nouns in WordNet. *Proceedings of EMNLP-2003*.
- Cruse, D. (1995). Polysemy and Related Phenomena from a Cognitive Linguistic Viewpoint. In : P. St-Dizier and É. Viegas (dir.) : *Computational Lexical Semantics*. Cambridge (G.-B.) : Cambridge University Press, 33-49.
- Delli Bovi C., J. Camacho-Collados, A. Raganato, and R. Navigli. (2017) Eurosense : Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of ACL*.
- Djemaa, M., Candito, M., Muller P. and Vieu L. (2016) Corpus annotation within the French Framenet : methodology and results, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), Portorož, Slovenia, 2016.
- Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A. and Biemann, C. (2016) A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *Proceedings of the LT4DH workshop at COLING 2016*, Osaka, Japan
- Fellbaum C. (1998) *WordNet : An Electronic Lexical Database*. Cambridge, MA : MIT Press
- Landes, S., Leacock, C., and Tengi, R. I. (1998). Building semantic concordances. In Fellbaum, C. (Ed.), WordNet : An Electronic Lexical Database, pp. 199–216. MIT Press.
- Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller (1990) WordNet : An online lexical database. *International Journal of Lexicography* 3(4). 235–244.

Références 2

- Navigli R. (2006) Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics* (COLING-ACL, Sydney, Australia). 105–112.
- Navigli R., Litkowski , K.C., and Hargraves, O. (2007) Semeval-2007 task 07 : Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval, Prague, Czech Republic). 30–35.
- Palmer M., Dang, H., and Fellbaum, C. (2007) Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *J. Nat. Lang. Eng.* 13, 2, 137–163
- Pedersen B. S, A. Braasch, A. Johannsen, H. Mart´nez Alonso, S. Nimb, S. Olsen, A. Søgaard, and N. Sørensen (2016) The semdax corpus–sense annotations with scalable sense inventories. In *LREC*.
- Petrolito T and F. Bond (2014). A survey of WordNet Annotated Corpora, In Proceedings of the Seventh Global WordNet Conference, Tartu, Estonia, pp. 236-243. Tartu, Estonia.
- Snyder B. and M. Palmer (2004) The English all-words task. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (Senseval-3, Barcelona, Spain). 41–43.
- Schneider N., B. Mohit, K. Oflazer and N. Smith (2012) Coarse lexical semantic annotation with supersenses : an Arabic case study. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju Island, Korea.
- Weischedel, Ralph, et al. OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia : Linguistic Data Consortium, 2013.