Exploitation et enrichissement de ressources lexicales crowdsourcées : WISIGOTH comme tentative de guidage des foules

M2 TAL - 12 Nov. 2012

Franck Sajous[*], Emmanuel Navarro[**], Bruno Gaume[*] et Yannick Chudy[*]
(*) CLLE-ERSS, CNRS & Univ. de Toulouse
(**) IRIT, CNRS & Univ. de Toulouse

Sommaire

1. Ressources lexicales
   - Contexte
   - Problème de l'évaluation
2. Crowdsourcing
   - Considérations générales
   - Utilisations en TAL
3. Wiktionary, *le compagnon lexical de Wikipédia*
   - Présentation
   - Croissance, nomenclature et couverture lexicale
4. WISIGOTH
   - Premières expériences de densification des réseaux
   - Enrichissement semi-automatique validé *par les foules*
   - Extension Firefox
   - Travaux connexes et perspectives

Contexte historique

We desperately need linguistic resources ! [Sekine, 2010]

- applications TAL : gourmandes en ressources lexicales
- anglais : WordNet ([Fellbaum, 1998])
- projets EuroWordNet ([Vossen, 1998]), puis BalkaNet ([Tufiș and Cristea, 2002])
- pour le français : pas de ressources et/ou ressources non libres et/ou payantes et/ou de piètre qualité
- partie française d'EuroWordNet : seulement noms et verbes, faible couverture, incohérences, ressource payante et figée

Tentatives pour combler ces lacunes

Vraies tentatives

- tentatives d'améliorations automatiques ([Jacquin et al., 2007]) intéressantes mais pas suffisantes
- pour certaines langues : ≪ juste ≫ décider de passer les ressources sous licence libre (e.g. DanNet [Pedersen, 2010])

Tentatives pour combler ces lacunes

Vraies tentatives

- tentatives d'améliorations automatiques ([Jacquin et al., 2007]) intéressantes mais pas suffisantes
- pour certaines langues : ≪ juste ≫ décider de passer les ressources sous licence libre (e.g. DanNet [Pedersen, 2010])

Phénomènes de mode

- créer de nouvelles normes de **métadonnées**
- créer de nouveaux **formats d'encodage**
- réfléchir à l'**interopérabilité** :
  concevoir des **services web** qui respectent les normes de métadonnées et d'encodage

Tentatives pour combler ces lacunes

Vraies tentatives

- tentatives d'améliorations automatiques ([Jacquin et al., 2007]) intéressantes mais pas suffisantes
- pour certaines langues : ≪ juste ≫ décider de passer les ressources sous licence libre (e.g. DanNet [Pedersen, 2010])

Phénomènes de mode

- créer de nouvelles normes de **métadonnées**
- créer de nouveaux **formats d'encodage**
- réfléchir à l'**interopérabilité** :
  concevoir des **services web** qui respectent les normes de métadonnées et d'encodage

oui, mais... quel contenu ?

Autres tentatives pour combler ces lacunes

Encourager (sic !) les ressources lexicales

→ financer des réseaux européens :

- Clarin (Common Language Resources and Technology Infrastructure)
- FLaReNet (Fostering Language Resources Network)

puis finalement nationaux :

- IR Corpus ?

Autres tentatives pour combler ces lacunes

Encourager (sic !) les ressources lexicales

→ financer des réseaux européens :

- Clarin (Common Language Resources and Technology Infrastructure)
- FLaReNet (Fostering Language Resources Network)

puis finalement nationaux :

- IR Corpus ?

qui. . .

- créent de nouvelles normes de métadonnées
- créent de nouveaux formats d'encodage
- réfléchissent à l'interopérabilité (créent des grilles de services web)
- encouragent les ressources lexicales

Autres tentatives pour combler ces lacunes

Encourager (sic !) les ressources lexicales

→ financer des réseaux européens :

- Clarin (Common Language Resources and Technology Infrastructure)
- FLaReNet (Fostering Language Resources Network)

puis finalement nationaux :

- IR Corpus ?

qui. . .

- créent de nouvelles normes de métadonnées
- créent de nouveaux formats d'encodage
- réfléchissent à l'interopérabilité (créent des grilles de services web)
- encouragent les ressources lexicales

oui, mais. . . quel contenu ?

Autres tentatives pour combler ces lacunes (2)

Construire automatiquement des ressources

Plusieurs méthodes :

- depuis l'extraction en corpus de relations par patrons lexico-syntaxiques : [Hearst, 1992] (hyperonymie)
- à divers calculs de similarité sémantique utilisant cooccurrents version sacs de mots ou contextes syntaxiques : [Curran and Moens, 2002, Van der Plas and Bouma, 2005, Heylen et al., 2008] (synonymie)
- en passant par la fertilisation mutuelles de ressources (projection de relations sémantiques suivant des liens de traduction) [Soria et al., 2009]. Pour le français : WOLF

Problème : quelle qualité ? Quelle évaluation pour estimer cette qualité ?

Quelle(s) évaluation(s) pour les ressources lexicales ?

Comparaison avec un étalon (Gold Standard)

Quelle(s) évaluation(s) pour les ressources lexicales ?

Comparaison avec un étalon (Gold Standard)

- si on a un déjà GS, pourquoi construire une autre ressource ?
- désaccord entre ressource évaluée et GS : quelle interprétation (GS non nécessairement exhaustif, éventuellement construit dans une optique particulière), rôle de la granularité

Quelle(s) évaluation(s) pour les ressources lexicales ?

Comparaison avec un étalon (Gold Standard)

- si on a un déjà GS, pourquoi construire une autre ressource ?
- désaccord entre ressource évaluée et GS : quelle interprétation (GS non nécessairement exhaustif, éventuellement construit dans une optique particulière), rôle de la granularité

Évaluation manuelle (d'un échantillon)

- quels juges, quel accord inter-juge ?
- [Murray and Green, 2004] : sur une tâche de WSD, accord élevé lorsque compétences lexicales similaires (pas lorsque compétences élevées)

Quelle(s) évaluation(s) pour les ressources lexicales ?

Comparaison avec un étalon (Gold Standard)

- si on a un déjà GS, pourquoi construire une autre ressource ?
- désaccord entre ressource évaluée et GS : quelle interprétation (GS non nécessairement exhaustif, éventuellement construit dans une optique particulière), rôle de la granularité

Évaluation manuelle (d'un échantillon)

- quels juges, quel accord inter-juge ?
- [Murray and Green, 2004] : sur une tâche de WSD, accord élevé lorsque compétences lexicales similaires (pas lorsque compétences élevées)

Évaluation par la tâche

- étude de l'impact de l'utilisation d'une ressource sur les résultats produits par un système.
- évaluer les résultats du système... revient à construire un GS !
  (voir [Kilgarriff, 1998] sur la préparation d'un GS pour SENSEVAL).
- importance de la granularité : [Palmer et al., 2007] accord sur tâche de WSD +10 à +20% en regroupant les sens d'un dictionnaire pour régler les désaccords ≪ à la marge ≫

Naissance (du mot)

Forgé par [Howe, 2006]

"Remember outsourcing ? Sending jobs to India and China is so 2003. The new pool of cheap labor : everyday people using their spare cycles to create content, solve problems, even do corporate R&D."
`http://www.wired.com/wired/archive/14.06/crowds.html`

Crowdsourcing : la traduction ≪ *approvisionnement par les foules* ≫ associée à l'idée de mutualisation de données/connaissances ne reflète pas la diversité des pratiques

Naissance (du mot)

Forgé par [Howe, 2006]

"Remember outsourcing ? Sending jobs to India and China is so 2003. The new pool of cheap labor : everyday people using their spare cycles to create content, solve problems, even do corporate R&D."
`http://www.wired.com/wired/archive/14.06/crowds.html`

Crowdsourcing : la traduction ≪ *approvisionnement par les foules* ≫ associée à l'idée de mutualisation de données/connaissances ne reflète pas la diversité des pratiques

Au moment de son apparition, le terme crowdsourcing, souvent traduit par ≪ approvisionnement par les foules ≫, semblait renvoyer à une notion assez floue, véhiculant une idée de partage et de mise en commun de ressources.
Le crowdsourcing, généralement associé à l'idée de mutualisation du savoir (ex. wikipedia) ou à celle de co-création ou de coproduction artistique, est en passe d'évoluer vers des pratiques beaucoup plus variées.

Utilisé par des groupements d'intérêt dans une visée citoyenne, mais aussi par des entreprises voire par des groupes industriels, le crowdsourcing recouvre aujourd'hui des enjeux économiques et sociaux très importants.

Taxonomie de [Quinn and Bederson, 2009]

Distributed human computation

"[...]we now define DHC for the purposes of this paper as 'systems of computers and large numbers of humans that work together in order to solve problems that could not be solved by either computers or humans alone"'.

- Plusieurs critères : motivation, qualité (e.g. détection de fraude), agrégation, compétences humaines, temps de participation, charge cognitive
- Types de DHC :
  - *Games with a prupose* (GWAP) ;
  - *Mechanized labor*, crowdsourcing qui implique une rétribution financière (e.g. AMT) ;
  - *Wisdom of Crowds*, ≪ le poids du panier garni ≫ ;
  - *Crowdsourcing*, aide bénévole de volontaires ;
  - *Dual-Purpose Work* (e.g. ReCAPTCHA) ;
  - *Grand search, Human based algorithms, knowledge from volunteer contributors,* etc.

Quelques exemples

- Wikipédia, évidemment. Projets satellites : Wiktionary, Wikisource, etc.

Quelques exemples

- Wikipédia, évidemment. Projets satellites : Wiktionary, Wikisource, etc.
- Wikitravel

Quelques exemples

- Wikipédia, évidemment. Projets satellites : Wiktionary, Wikisource, etc.
- Wikitravel
- tripadvisor

  `http://www.rue89.com/rue89-eco/2012/10/27/`

  `faux-avis-comment-les-pros-dupent-les-internautes-236572`

Quelques exemples

- Wikipédia, évidemment. Projets satellites : Wiktionary, Wikisource, etc.
- Wikitravel
- tripadvisor

  `http://www.rue89.com/rue89-eco/2012/10/27/`

  `faux-avis-comment-les-pros-dupent-les-internautes-236572`
- Openstreetmap

Quelques exemples

- Wikipédia, évidemment. Projets satellites : Wiktionary, Wikisource, etc.
- Wikitravel
- tripadvisor

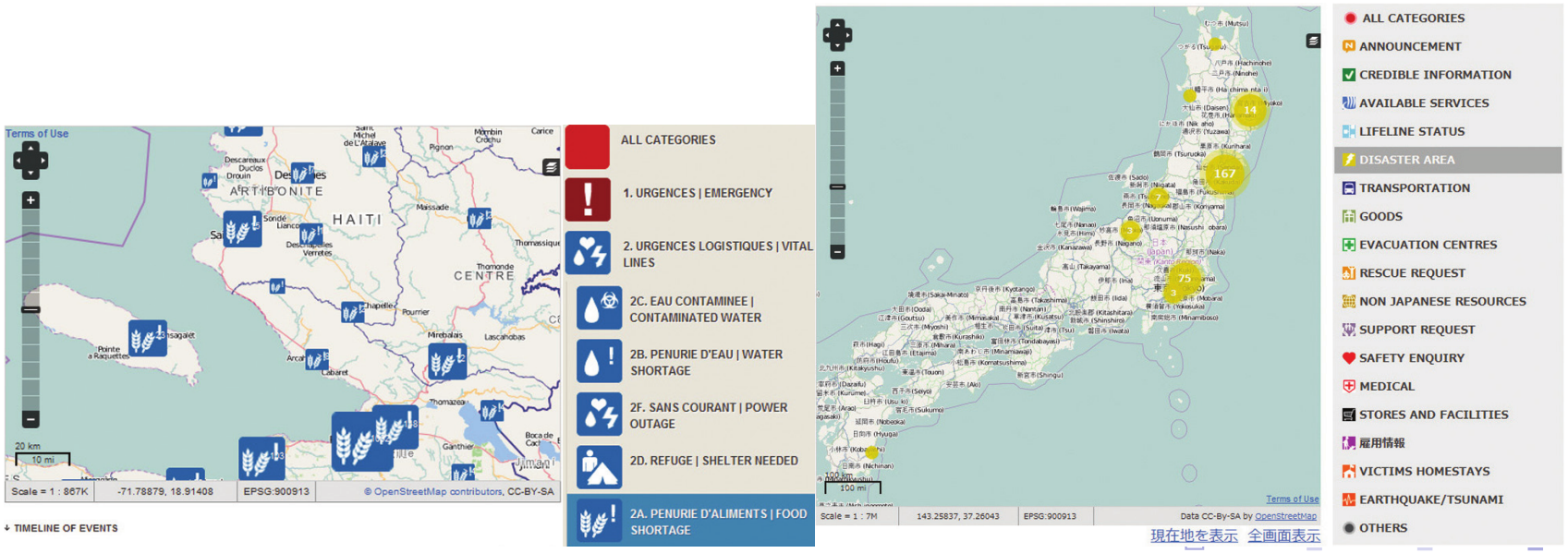
  `http://www.rue89.com/rue89-eco/2012/10/27/`

  `faux-avis-comment-les-pros-dupent-les-internautes-236572`
- Openstreetmap
- `ushahidi.com` (qui repose sur Openstreetmap) : plateforme de gestion de crise lors de catastrophes naturelles et crises sanitaires ([Gao et al., 2011])

Quelques exemples

- Wikipédia, évidemment. Projets satellites : Wiktionary, Wikisource, etc.
- Wikitravel
- tripadvisor

  `http://www.rue89.com/rue89-eco/2012/10/27/faux-avis-comment-les-pros-dupent-les-internautes-236572`
- Openstreetmap
- `ushahidi.com` (qui repose sur Openstreetmap) : plateforme de gestion de crise lors de catastrophes naturelles et crises sanitaires ([Gao et al., 2011])



Taxonomie (ternaire, un peu simpliste) de [Sajous et al., 2012]

...qui n'avaient pas lu [Quinn and Bederson, 2009] à l'époque (en 2010)

- GWAP : *"Labeling images with a computer game"* ([von Ahn and Dabbish, 2004]), puis en TAL : Jeux de mots ([Lafourcade, 2007]), Phrase Detective ([Chamberlain et al., 2009])
- AMT : annotation par des naïfs pour apprentissage supervisé ([Snow et al., 2008])
- Piggybacking : WISIGOTH [Sajous et al., 2010]

Note : *crowdsourcing/wisdom of crowds* pas utilisés directement, mais comme source de données primaire, e.g. calcul de similarité sémantique ([Zesch et al., 2008]).

Critiques

- qualité des contenus crowdsourcés : polémique [Giles, 2005] vs. Britanica. Pas très intéressante en 2005, obsolète aujourd'hui
- éthique d'AMT...

Éditions de langues

`http://www.wiktionary.org`

- *Wiktionary* : désigne à la fois ≪ le projet ≫ (l'ensemble des éditions de langues) et l'édition anglaise
- *Wiktionnaire* : édition française de Wiktionary

Wiktionnaire

≪ Le Wiktionnaire est un dictionnaire universel libre en développement fondé sur les contributions coopératives des internautes dans le cadre d'un wiki utilisant la technologie du Web 2.0.
Le Wiktionnaire (appellation francisée) est la partie francophone du projet multilingue Wiktionary [. . .]
Son objectif est seulement descriptif : il ne s'agit ni de défendre le français ou une autre langue, ni d'être normatif. Il ne juge donc pas la valeur des mots et n'essaie pas de leur donner ou de leur refuser son aval. ≫

`http://fr.wiktionary.org/wiki/Wiktionnaire:A_propos`

Structure d'un article

Contenu des articles

boot

Etymology 1

Middle English, from Old French *bote*

Noun

1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms

* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations

shoe

* French: botte
* Spanish : bota

kick - see kick

Verb

1. To kick
   *I **booted** the ball*
2. To disconnect
   *I got **booted** from the chatroom*

Synonyms

* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2

Akin to Old Norse *bót*

Structure d'un article

Contenu des articles

- étymologie

boot

Etymology 1 *etymology*
Middle English, from Old French *bote*

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations
shoe
* French: botte
* Spanish : bota

kick - see kick

Verb
1. To kick
   *I **booted** the ball*
2. To disconnect
   *I got **booted** from the chatroom*

Synonyms
* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2
Akin to Old Norse *bót*

Structure d'un article

Contenu des articles

- étymologie
- parties du discours

boot

Etymology 1 *etymology*
Middle English, from Old French *bote*

Noun *part of speech (noun)*
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations
shoe
* French: botte
* Spanish : bota

kick - see kick

Verb *part of speech (verb)*
1. To kick
   *I **booted** the ball*
2. To disconnect
   *I got **booted** from the chatroom*

Synonyms
* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2
Akin to Old Norse *bót*

Structure d'un article

Contenu des articles

- étymologie
- parties du discours
  - définitions, examples

boot

Etymology 1 *etymology*
Middle English, from Old French *bote*

Noun *part of speech (noun)*
1. A heavy shoe *wordsenses*
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations
shoe
* French: botte
* Spanish : bota

kick - see kick

Verb *part of speech (verb)*
1. To kick *wordsense #1*
*I **booted** the ball*
2. To disconnect *wordsense #2*
*I got **booted** from the chatroom*

Synonyms
* (*kick*): kick
* (disconnect): kick

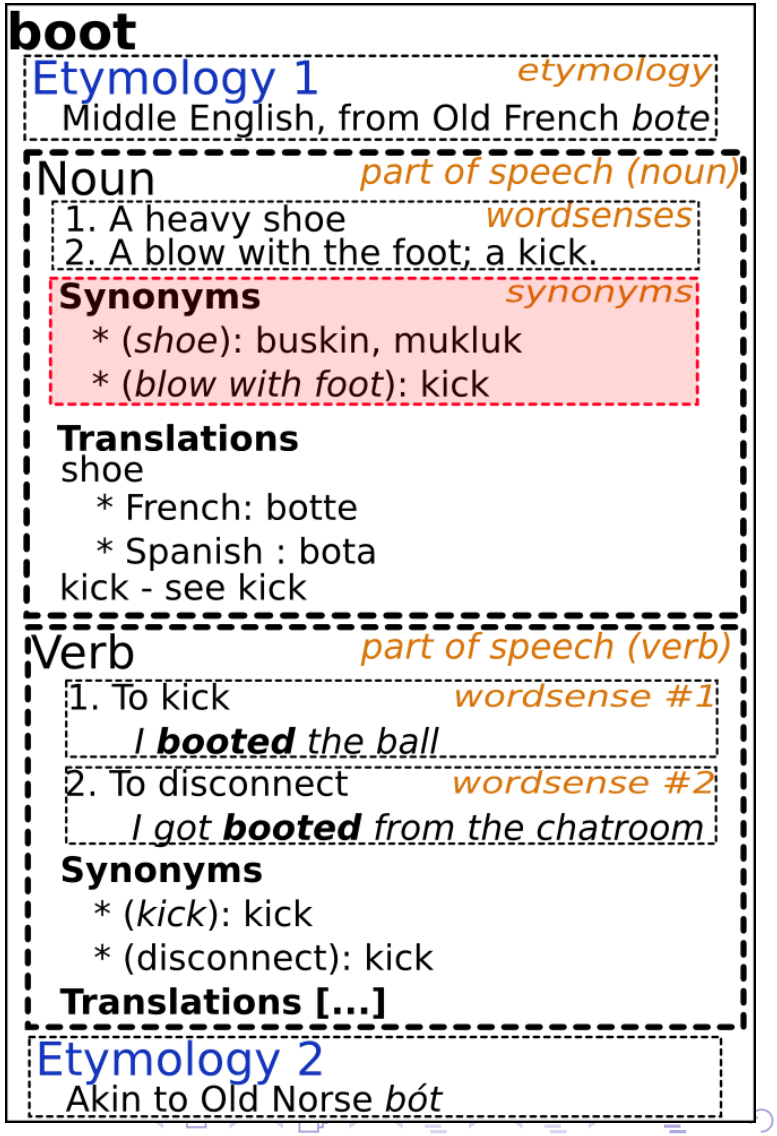
Translations [...]

Etymology 2
Akin to Old Norse *bót*

Structure d'un article

Contenu des articles

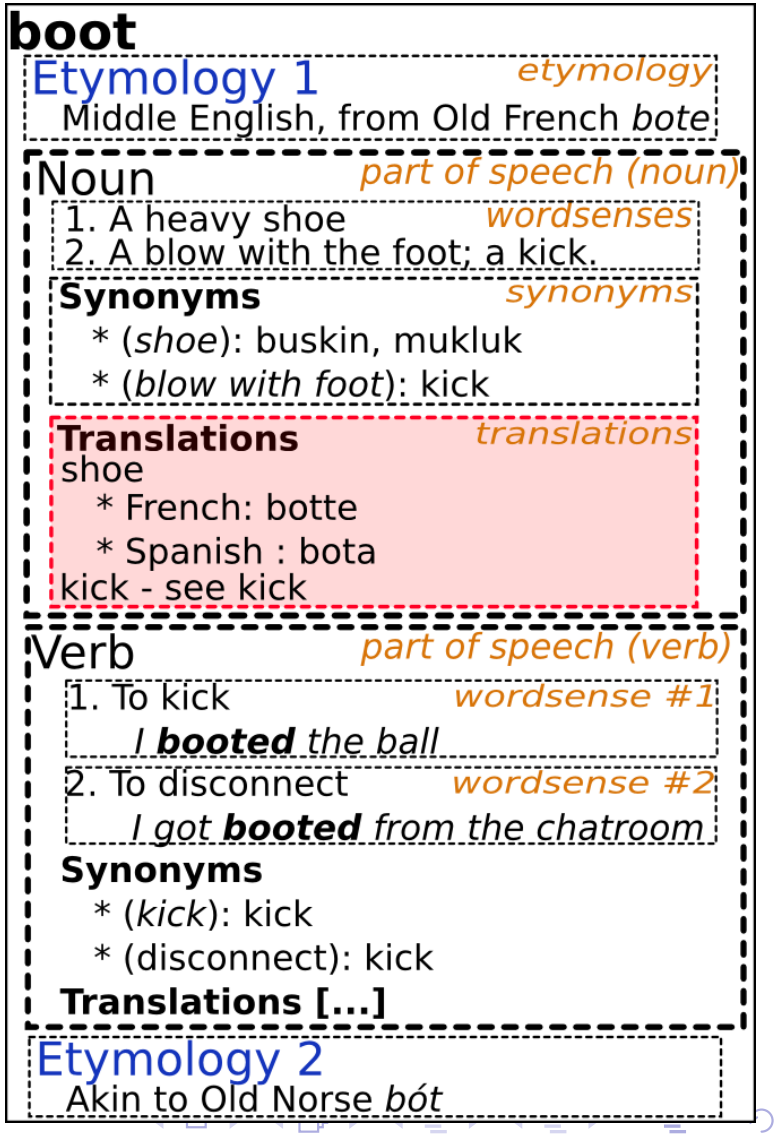
- étymologie
- parties du discours
  - définitions, examples
  - relations sémantiques



Structure d'un article

Contenu des articles

- étymologie
- parties du discours
  - définitions, examples
  - relations sémantiques
  - traductions



Structure d'un article

Contenu des articles

- étymologie
- parties du discours
  - définitions, examples
  - relations sémantiques
  - traductions

Le cas "régulier", mais...

- contenu & structure hétérogènes d'un langage à l'autre et même au sein d'un même langage (déviances par rapport à un standard qui n'existe pas).
- $X \in syns(Y) \nRightarrow Y \in syns(X)$, $X \in trads(Y) \nRightarrow Y \in trads(X)$, etc.
- parfois (e.g. en italien), une partie relation sém (e.g. syn) correspond à tous les POS.
- gestion des sous-sens... fantaisiste.

boot

Etymology 1

Middle English, from Old French *bote*

Noun

1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms

* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Translations

shoe
* French: botte
* Spanish : bota

kick - see kick

Verb

1. To kick
   *I **booted** the ball*
2. To disconnect
   *I got **booted** from the chatroom*

Synonyms

* (*kick*): kick
* (disconnect): kick

Translations [...]

Etymology 2

Akin to Old Norse *bót*

Encodage des données : *wikicode* (ou *wikitexte*)

```
==English==
===Etymology 1===
{{etyl|enm}} {{term|boot|lang=enm||shoe}}, [...]
====Noun====
# A heavy [[shoe]] that [[cover]]s part of
the leg.
# A [[blow]] with the foot; a [[kick]].
=====Synonyms=====
* {{sense|shoe}} [[buskin]], [[mukluk]]
* {{sense|blow with foot}} [[kick]]
=====Translations=====
* French: [[botte]] {{f}}, [[bottine]] {{f}}
* Spanish: {{t+|es|bota|f}}
{{trans-see|kick}}

====Verb====
[...]
```

Date	Code wiki
2004-03-22	===Nom===
2004-04-01	===[[Nom]]===
2004-04-01	===[[Nom commun]]===
2004-09-18	=== [[Nom commun]] ===
2004-10-04	=== Nom commun ===
2004-11-02	{{-nom-}}
2006-09-29	{{-nom-|fr}}

- Pas de définition finie de la syntaxe de MediaWiki (système de gestion de contenu). → pas de vérification possible lors d'une modification
- Syntaxe réellement plus simple qu'HTML ?
- Syntaxe spécifique à une édition de langue donnée.
- Syntaxe en constante évolution

Date	Code wiki
2004-03-22	====Traductions====
2004-04-01	====[[Traduction]]s====
2004-10-04	==== Traductions ====
2004-11-01	{{-trans-}}
2005-04-13	{{-trad-}}
2009-03-12	{{-trad-trier}}

Traductions, mars 2004 :

```
* albanien : [[fjalor]]
* [[breton]] : [[geriadur]]
```

Traductions, octobre 2004 :

```
* {{sq}} : [[fjalor]]
* {{br}} : [[geriadur]]
```

Ancrage des relations

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

Ancrage des relations

Ancrage des relations

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* buskin, mukluk
* kick

?

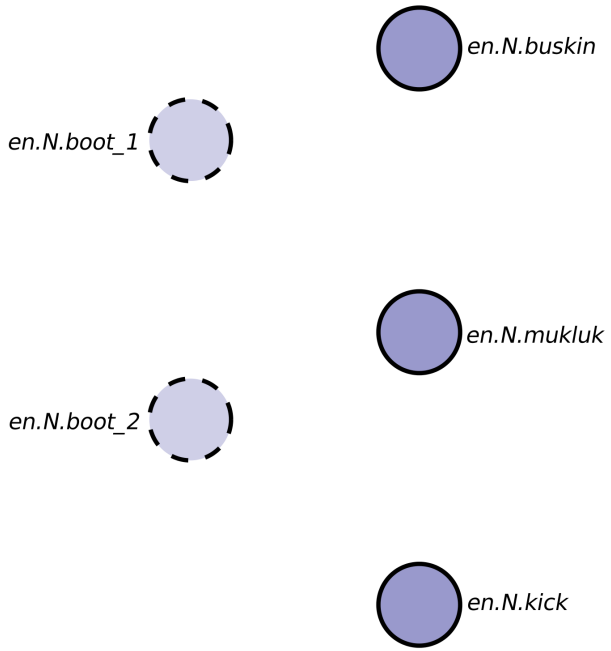
Ancrage des relations

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* buskin, mukluk
* kick

?



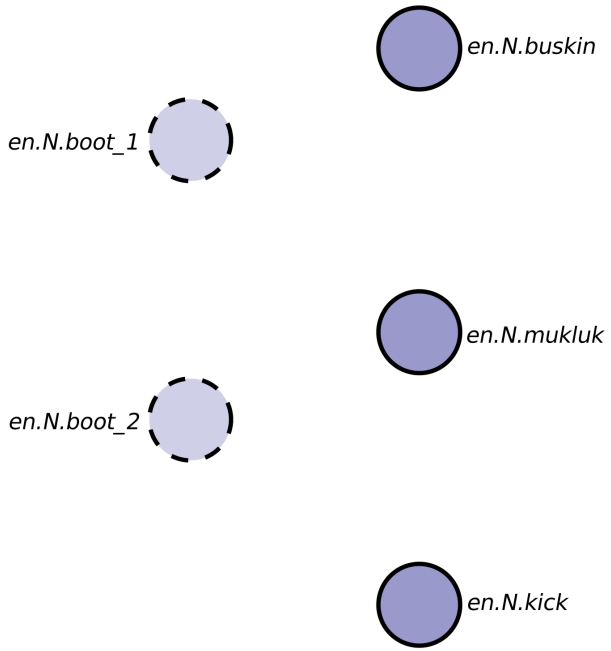
Ancrage des relations

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

?



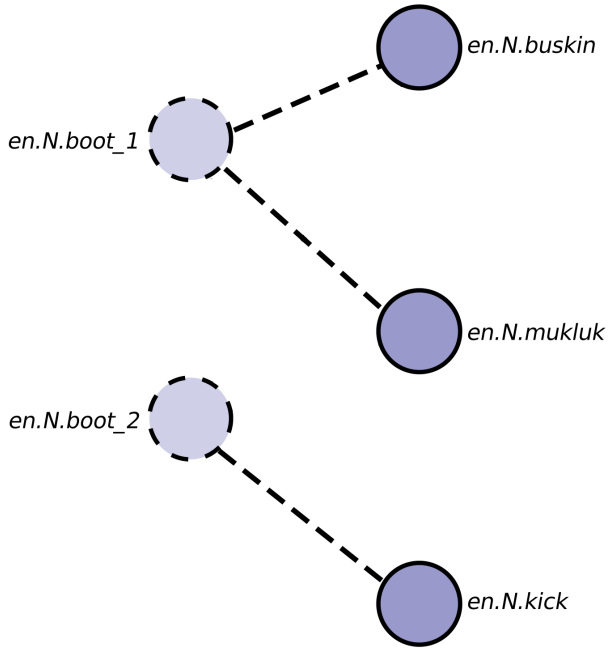
Ancrage des relations

boot
English

Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

?

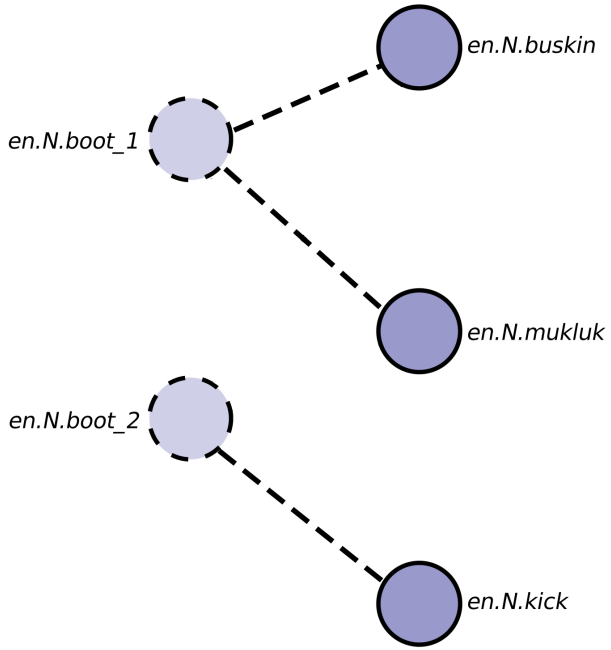


Ancrage des relations

boot
English
Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

?



buskin (N)

1. A half-boot
2. A type of boot worn by the ancient Athenian tragic actors

mukluk (N)

1. A soft boot made of reindeer skin or sealskin and worn by Inuit.

kick (N)

1. A hit or strike with the leg or foot
2. The action of swinging a foot or leg
3. Sth that tickles the fancy
4. (Internet) The removal of a person from an online activity
5. (figuratively) Any bucking motion of an object that lacks legs or feet

Ancrage des relations

boot
English
Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

?

buskin (N)

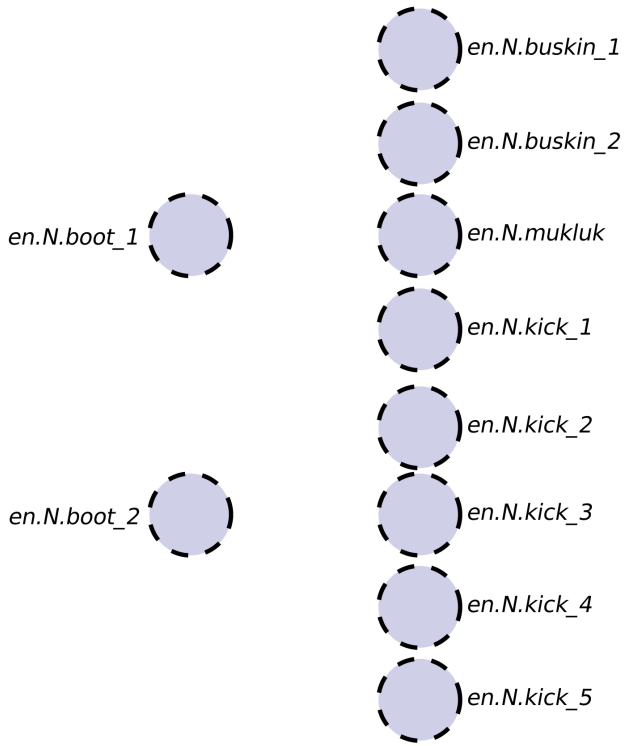
1. A half-boot
2. A type of boot worn by the ancient Athenian tragic actors

mukluk (N)

1. A soft boot made of reindeer skin or sealskin and worn by Inuit.

kick (N)

1. A hit or strike with the leg or foot
2. The action of swinging a foot or leg
3. Sth that tickles the fancy
4. (Internet) The removal of a person from an online activity
5. (figuratively) Any bucking motion of an object that lacks legs or feet

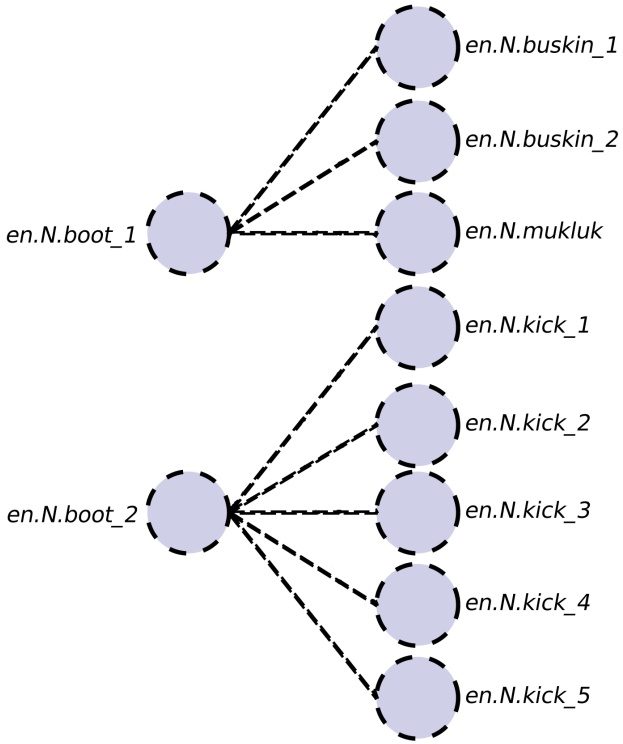


Ancrage des relations

boot
English
Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

?



buskin (N)

1. A half-boot
2. A type of boot worn by the ancient Athenian tragic actors

mukluk (N)

1. A soft boot made of reindeer skin or sealskin and worn by Inuit.

kick (N)

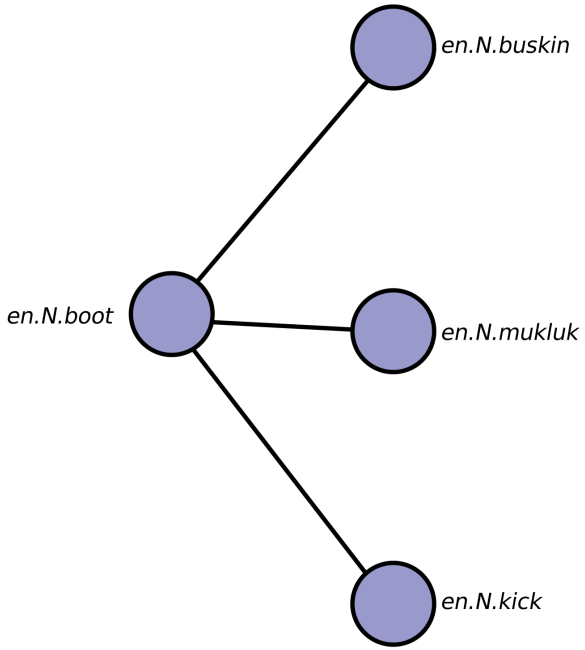
1. A hit or strike with the leg or foot
2. The action of swinging a foot or leg
3. Sth that tickles the fancy
4. (Internet) The removal of a person from an online activity
5. (figuratively) Any bucking motion of an object that lacks legs or feet

Ancrage des relations

boot
English
Noun
1. A heavy shoe
2. A blow with the foot; a kick.

Synonyms
* (*shoe*): buskin, mukluk
* (*blow with foot*): kick

?



buskin (N)

1. A half-boot
2. A type of boot worn by the ancient Athenian tragic actors

mukluk (N)

1. A soft boot made of reindeer skin or sealskin and worn by Inuit.

kick (N)

1. A hit or strike with the leg or foot
2. The action of swinging a foot or leg
3. Sth that tickles the fancy
4. (Internet) The removal of a person from an online activity
5. (figuratively) Any bucking motion of an object that lacks legs or feet

Ancrage des relations, autres éditions de langue

Nom commun

papillon /pa.pi.jɔ̃/ *masculin*

1. *(Zoologie)* Insecte qui a quatre ailes, couvertes d'écailles fines.
   - *Le ver le secrète lorsqu'il est arrivé au moment où il doit se mé*
     filature; 1re partie: Fibres animales & minérales, Encyclopédie Roret, 1
   - ***Papillon** blanc, rouge, bigarré, etc. — Les enfants courent apre*
2. *(Figuré)* *(Familier)* Personne qui, se laissant tromper par des appar
   piège.
   - *Il va se brûler à la chandelle comme un **papillon**.*
3. *(Figuré)* *(Familier)* Esprit léger qui voltige d'objets en objets.
   - *C'est un **papillon**.*
4. *(Figuré)* *(Familier)* Bagatelle, considération futile.
   - *Courir après les **papillons**.*
5. *(Par analogie)* Petit papier détachable à bande semi-adhésive ou s
   revue, pour donner un avis au lecteur, indiquer un erratum, etc.
   - *Mettre un **papillon**.*
6. *(Familier)* Procès verbal de contravention.
   - *Charlotte lui désigne un **papillon** de contravention sur le pare-*
     nouvelles, Éditeur L'Âge d'Homme, 1991).
7. *(Mécanique)* Volet orientable posé dans un conduit et permettant
   - *La pédale (auto) ou manette (moto) d'accélérateur n'agit que s*

Synonymes

Insecte :

- lépidoptère *(Science)* (1)
- prune, PV (9)

Nage :

- papillon-dauphin
- brasse papillon

Ancrage des relations, autres éditions de langue

Nom commun

papillon /pa.pi.jɔ̃/ *masculin*

1. *(Zoologie)* Insecte qui a quatre ailes, couvertes d'écailles fines.
   - *Le ver le secrète lorsqu'il est arrivé au moment où il doit se mé*
     filature; 1re partie: Fibres animales & minérales, Encyclopédie Roret, 1
   - ***Papillon** blanc, rouge, bigarré, etc. — Les enfants courent apre*
2. *(Figuré)* *(Familier)* Personne qui, se laissant tromper par des appa
   piège.
   - *Il va se brûler à la chandelle comme un **papillon**.*
3. *(Figuré)* *(Familier)* Esprit léger qui voltige d'objets en objets.
   - *C'est un **papillon**.*
4. *(Figuré)* *(Familier)* Bagatelle, considération futile.
   - *Courir après les **papillons**.*
5. *(Par analogie)* Petit papier détachable à bande semi-adhésive ou s
   revue, pour donner un avis au lecteur, indiquer un erratum, etc.
   - *Mettre un **papillon**.*
6. *(Familier)* Procès verbal de contravention.
   - *Charlotte lui désigne un **papillon** de contravention sur le pare-*
     nouvelles, Éditeur L'Âge d'Homme, 1991).
7. *(Mécanique)* Volet orientable posé dans un conduit et permettant
   - *La pédale (auto) ou manette (moto) d'accélérateur n'agit que s*

Synonymes

Insecte :

- lépidoptère *(Science)* (1)
- prune, PV (9)

Nage :

- papillon-dauphin
- brasse papillon

Noun

butterfly (*plural* **butterflies**)

1. A flying insect of the order *Lepidoptera*, disting
2. (*now rare*) Someone seen as being unserious
3. The butterfly stroke. [from 20th c.]
4. A use of surgical tape, cut into thin strips

Synonyms

- lep

Ancrage des relations, autres éditions de langue

Nom commun

papillon /pa.pi.jɔ̃/ *masculin*

1. *(Zoologie)* Insecte qui a quatre ailes, couvertes d'écailles fines.
   - *Le ver le secrète lorsqu'il est arrivé au moment où il doit se mé*
     filature; 1re partie: Fibres animales & minérales, Encyclopédie Roret, 1
   - ***Papillon** blanc, rouge, bigarré, etc. — Les enfants courent apre*
2. *(Figuré)* *(Familier)* Personne qui, se laissant tromper par des appar
   piège.
   - *Il va se brûler à la chandelle comme un **papillon**.*
3. *(Figuré)* *(Familier)* Esprit léger qui voltige d'objets en objets.
   - *C'est un **papillon**.*
4. *(Figuré)* *(Familier)* Bagatelle, considération futile.
   - *Courir après les **papillons**.*
5. *(Par analogie)* Petit papier détachable à bande semi-adhésive ou s
   revue, pour donner un avis au lecteur, indiquer un erratum, etc.
   - *Mettre un **papillon**.*
6. *(Familier)* Procès verbal de contravention.
   - *Charlotte lui désigne un **papillon** de contravention sur le pare-*
     nouvelles, Éditeur L'Âge d'Homme, 1991).
7. *(Mécanique)* Volet orientable posé dans un conduit et permettant
   - *La pédale (auto) ou manette (moto) d'accélérateur n'agit que s*

Synonymes

Insecte :

- lépidoptère *(Science)* (1)
- prune, PV (9)

Nage :

- papillon-dauphin
- brasse papillon

Noun

butterfly (*plural* **butterflies**)

1. A flying insect of the order *Lepidoptera*, disting
2. (*now rare*) Someone seen as being unserious
3. The butterfly stroke. [from 20th c.]
4. A use of surgical tape, cut into thin strips

Synonyms

- lep

Sostantivo

farfalla (approfondimento) *f sing* (p

1. (*entomologia*) insetto dell'ordine d

Sillabazione

far | fàl | la

Pronuncia

IPA: /far'falla/

Iperonimi

- insetto

Ancrage des relations, autres éditions de langue

Nom commun

papillon /pa.pi.jɔ̃/ *masculin*

1. *(Zoologie)* Insecte qui a quatre ailes, couvertes d'écailles fines.
   - *Le ver le secrète lorsqu'il est arrivé au moment où il doit se mé*
     filature; 1re partie: Fibres animales & minérales, Encyclopédie Roret, 1
   - ***Papillon** blanc, rouge, bigarré, etc. — Les enfants courent apré*
2. *(Figuré)* *(Familier)* Personne qui, se laissant tromper par des appar
   piège.
   - *Il va se brûler à la chandelle comme un **papillon**.*
3. *(Figuré)* *(Familier)* Esprit léger qui voltige d'objets en objets.
   - *C'est un **papillon**.*
4. *(Figuré)* *(Familier)* Bagatelle, considération futile.
   - *Courir après les **papillons**.*
5. *(Par analogie)* Petit papier détachable à bande semi-adhésive ou s
   revue, pour donner un avis au lecteur, indiquer un erratum, etc.
   - *Mettre un **papillon**.*
6. *(Familier)* Procès verbal de contravention.
   - *Charlotte lui désigne un **papillon** de contravention sur le pare-*
     nouvelles, Éditeur L'Âge d'Homme, 1991).
7. *(Mécanique)* Volet orientable posé dans un conduit et permettant
   - *La pédale (auto) ou manette (moto) d'accélérateur n'agit que s*

Synonymes

Insecte :

- lépidoptère *(Science)* (1)
- prune, PV (9)

Nage :

- papillon-dauphin
- brasse papillon

Noun

butterfly (*plural* **butterflies**)

1. A flying insect of the order *Lepidoptera*, disting
2. (*now rare*) Someone seen as being unserious
3. The butterfly stroke. [from 20th c.]
4. A use of surgical tape, cut into thin strips

Synonyms

- lep

Sostantivo

farfalla (approfondimento) *f sing* (p

1. (*entomologia*) insetto dell'ordine d

Sillabazione

far | fàl | la

Pronuncia

IPA: /far'falla/

Iperonimi

- insetto

papallona

Català

1. insecte lepidòpter

- efecte papallona

Ancrage des relations, autres éditions de langue

Nom commun

papillon /pa.pi.jɔ̃/ *masculin*

1. *(Zoologie)* Insecte qui a quatre ailes, couvertes d'écailles fines.
   - *Le ver le secrète lorsqu'il est arrivé au moment où il doit se mé*
     filature; 1re partie: Fibres animales & minérales, Encyclopédie Roret, 1
   - ***Papillon** blanc, rouge, bigarré, etc. — Les enfants courent apré*
2. *(Figuré)* *(Familier)* Personne qui, se laissant tromper par des appa
   piège.
   - *Il va se brûler à la chandelle comme un **papillon**.*
3. *(Figuré)* *(Familier)* Esprit léger qui voltige d'objets en objets.
   - *C'est un **papillon**.*
4. *(Figuré)* *(Familier)* Bagatelle, considération futile.
   - *Courir après les **papillons**.*
5. *(Par analogie)* Petit papier détachable à bande semi-adhésive ou s
   revue, pour donner un avis au lecteur, indiquer un erratum, etc.
   - *Mettre un **papillon**.*
6. *(Familier)* Procès verbal de contravention.
   - *Charlotte lui désigne un **papillon** de contravention sur le pare-*
     nouvelles, Éditeur L'Âge d'Homme, 1991).
7. *(Mécanique)* Volet orientable posé dans un conduit et permettant
   - *La pédale (auto) ou manette (moto) d'accélérateur n'agit que s*

Synonymes

Insecte :

- lépidoptère *(Science)* (1)
- prune, PV (9)

Nage :

- papillon-dauphin
- brasse papillon

Noun

butterfly (*plural* **butterflies**)

1. A flying insect of the order *Lepidoptera*, disting
2. (*now rare*) Someone seen as being unserious
3. The butterfly stroke. [from 20th c.]
4. A use of surgical tape, cut into thin strips

Synonyms

- lep

Schmetterling (Deutsch) [Bearbeiten]

Substantiv, *m* [Bearbeiten]

Worttrennung:
Schmet·ter·ling, Plural: Schmet·ter·lin·ge

Aussprache:
IPA: ['ʃmɛtɐlɪŋ], Plural: ['ʃmɛtɐˌlɪŋə]
Hörbeispiele: —, Plural: —

Bedeutungen:
[1] *Biologie:* Insekt mit farbig beschuppten Flügeln
[2] *Biologie:* im engeren Sinne ein Tagfalter (tagakt
[3] *Sport:* ein Schwimmstil
[4] *Segeln:* eine Stellung der Segel
[5] *Sport:* freier Salto mit halber Drehung am höchs

Herkunft:
von ostmd. Schmetten (von tschech. smetana (= s
stahlen Sahne und andere Milchprodukte.[1]

Synonyme:
[1] wissenschaftlich: *Lepidoptera*
[2] Tagfalter, Sommervogel
[1, 2] Falter

Sostantivo

farfalla (approfondimento) *f sing* (p

1. (*entomologia*) insetto dell'ordine d

Sillabazione

far | fàl | la

Pronuncia

IPA: /far'falla/

Iperonimi

- insetto

papallona

Català

1. insecte lepidòpter

- efecte papallona

Le mirage des nombres

En 2010, Wiktionary fait état de « *2 038 436 entries with English definitions from over 400 languages* »
Wiktionnaire : « *1 821 097 articles [qui] décrivent en français les mots de plus de 700 langues* ».
Ces « articles » incluent :

- les formes fléchies des noms, adjectifs et verbes ;
- des locutions exotiques ;
- des méta-articles ;
- la déscription de mots d'autres langues (que l'édition de langue considérée) ;
- les pages de discussion.

Le mirage des nombres

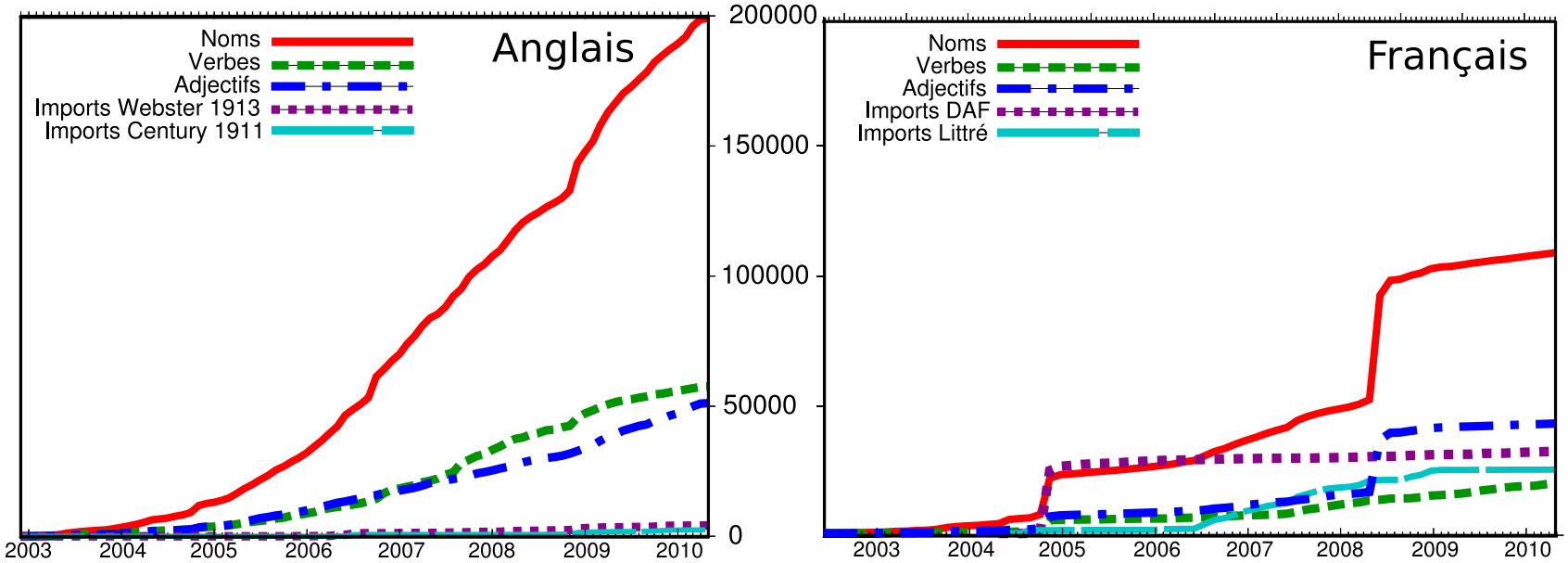
En 2010, Wiktionary fait état de « *2 038 436 entries with English definitions from over 400 languages* »
Wiktionnaire : « *1 821 097 articles [qui] décrivent en français les mots de plus de 700 langues* ».
Ces « articles » incluent :

- les formes fléchies des noms, adjectifs et verbes ;
- des locutions exotiques ;
- des méta-articles ;
- la déscription de mots d'autres langues (que l'édition de langue considérée) ;
- les pages de discussion.

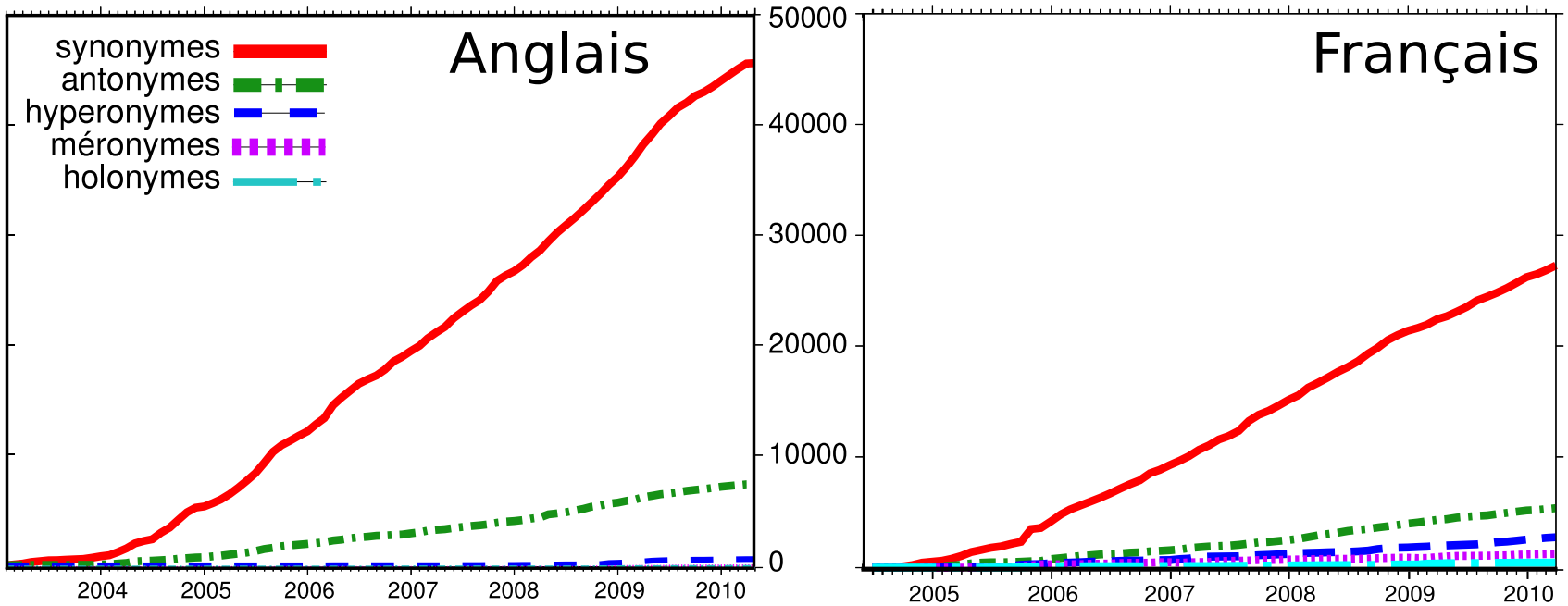
		2007			2010		
		N	**V**	**Adj**	**N**	**V**	**Adj**
FR	Vocables	38 973	6 968	11 787	106 068 **(x2.7)**	17 782 **(x2.6)**	41 725 **(x3.5)**
	Syns	9 670	1 793	2 522	17 054 **(x1.8)**	3 158 **(x1.8)**	4 111 **(x1.6)**
	Trads	106 061	43 319	25 066	153 060 **(x1.4)**	49 859 **(x1.2)**	32 949 **(x1.3)**
EN	Vocables	65 078	10 453	17 340	196 790 **(x3.0)**	67 649 **(x6.5)**	48 930 **(x2.8)**
	Syns	12 271	3 621	4 483	28 193 **(x2.3)**	8 602 **(x2.4)**	9 574 **(x2.1)**
	Trads	172 158	37 405	34 338	277 453 **(x1.6)**	70 271 **(x1.9)**	54 789 **(x1.6)**

Évolution historique - nombre de vocables

Dump historique du Wiktionnaire : après élimination des méta-articles mots étrangers, 1009144 pages, 3576223 révisions.

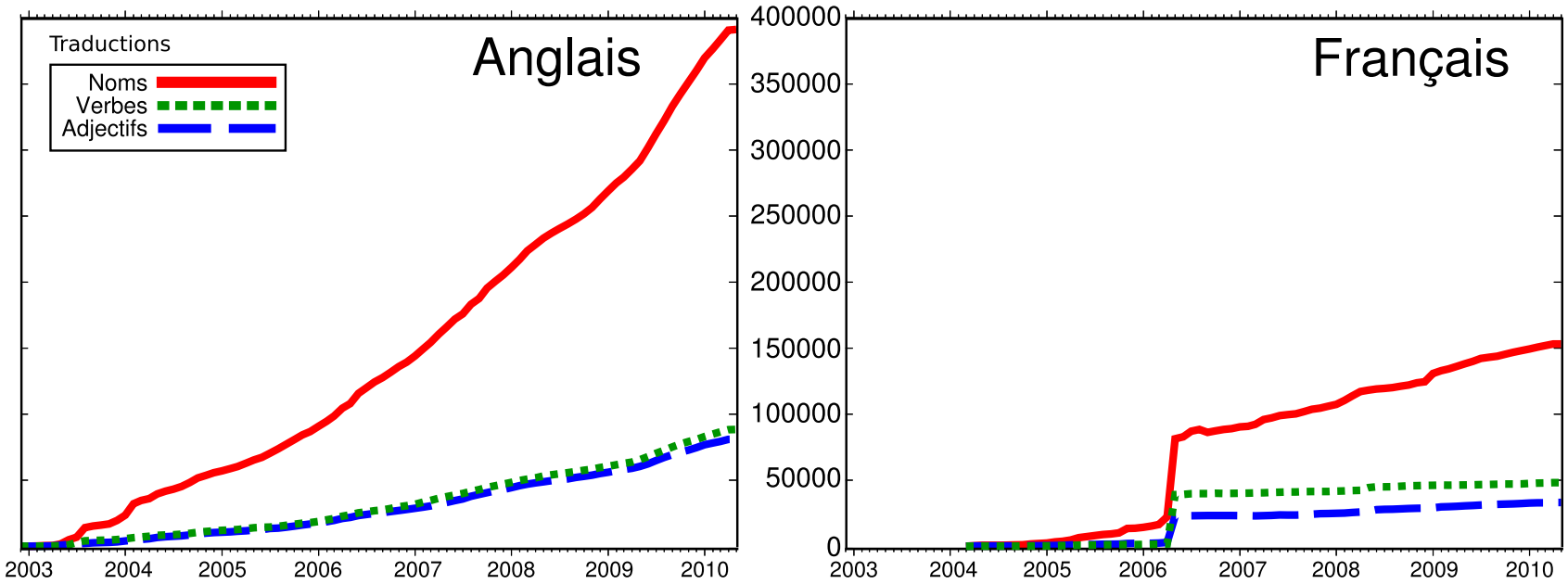


Évolution historique - nombre de relations sémantiques



(Toutes catégories syntaxiques confondues)

Évolution historique - nombre de traductions



Taille de nomenclature vs. couverture lexicale « effective »

Nomenclature Wiktionnaire vs. Morphalou.

	Taille de la nomenclature		
	Morphalou	Wiktionnaire	Intersection
N	41005	134203	29604
V	7384	18830	6964
Adj	15208	42263	10014

Taille de nomenclature vs. couverture lexicale « effective »

Nomenclature Wiktionnaire vs. Morphalou.

	Taille de la nomenclature		
	Morphalou	Wiktionnaire	Intersection
N	41005	134203	29604
V	7384	18830	6964
Adj	15208	42263	10014

Projection sur corpus

	Couverture lexiques/corpus (%)								
	Frantext 20^{e}			**Le Monde (1991-2000)**			**Wikipédia (2008)**		
	Morph.	Wikt.	M∪W	Morph.	Wikt.	M∪W	Morph.	Wikt.	M∪W
N	76,4	80,6	84,4	47,3	54,1	58,1	23,5	26,7	31,6
V	84,2	86,5	87,1	75,1	80,0	80,8	66,3	71,5	72,2
Adj	88,9	84,6	94,0	78,9	76,8	88,1	73,9	72,4	84,7

Taille de nomenclature vs. couverture lexicale « effective »

Nomenclature Wiktionnaire vs. Morphalou.

	Taille de la nomenclature		
	Morphalou	Wiktionnaire	Intersection
N	41005	134203	29604
V	7384	18830	6964
Adj	15208	42263	10014

Projection sur corpus

	Couverture lexiques/corpus (%)								
	Frantext 20^{e}			**Le Monde (1991-2000)**			**Wikipédia (2008)**		
	Morph.	Wikt.	M∪W	Morph.	Wikt.	M∪W	Morph.	Wikt.	M∪W
N	76,4	80,6	84,4	47,3	54,1	58,1	23,5	26,7	31,6
V	84,2	86,5	87,1	75,1	80,0	80,8	66,3	71,5	72,2
Adj	88,9	84,6	94,0	78,9	76,8	88,1	73,9	72,4	84,7

- Couverture /Frantext > couverture /LM > couverture /Wikipédia
- Wiktionnary couvre + de N et V (2 à 7%)
- Morphalou couvre + d'Adj (1 à 4%)
- Union : +5% pour les noms et +10% pour les adjs.

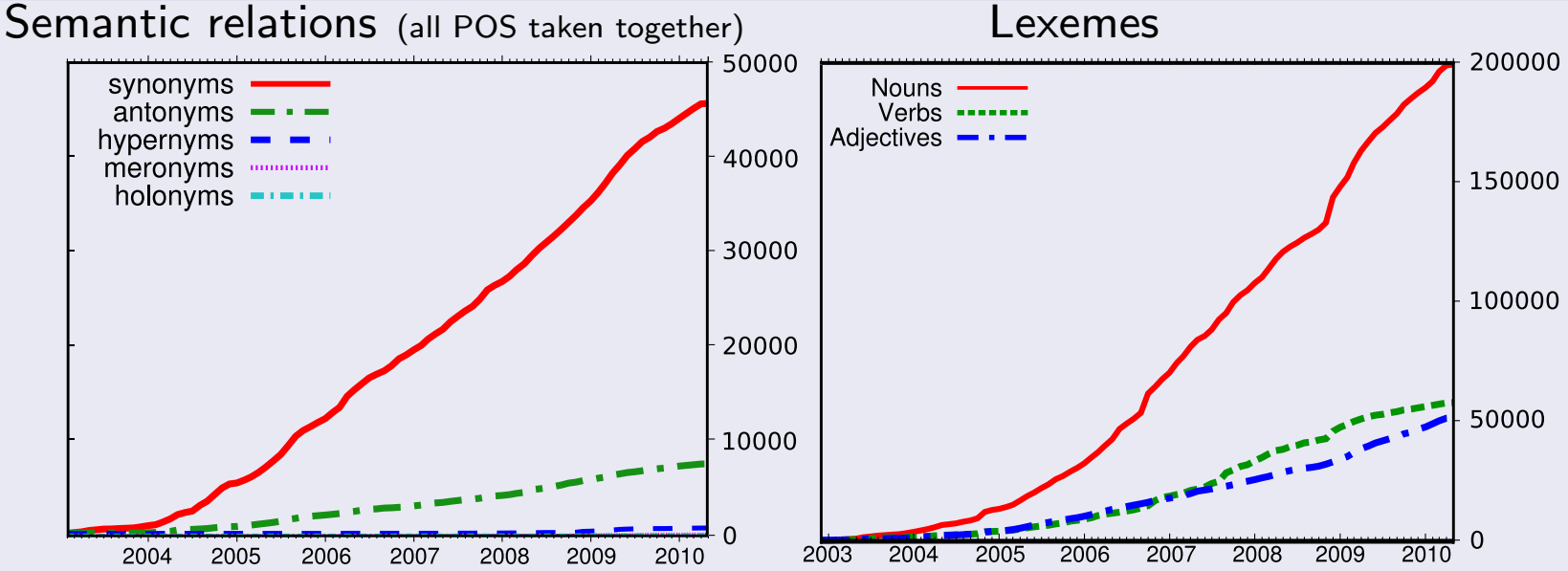
Évaluation de la couverture lexicale « effective »

Constats

- Confirmation des observations de [Zesch, 2010] : ne pas nécessairement mettre en compétition *wisdom of crowds* vs. *wisdom of linguists*, mais considérer leur complémentarité.
- Observation qualitative : pas seulement des néologismes issus de la sphère techno/internet (*googler, wikifier*), mais :
  - variations diatopiques : *dracher*, *diplomerie*
  - termes spécialisés : *clitique*, *métier* (adj)
  - usage courant : *sinogramme, homophobie, sociétal, fractal, ergonomique, médicaliser, étanchéifier, désactiver, décélérer, paramétrer*

Croissance vocables vs. relations sémantiques

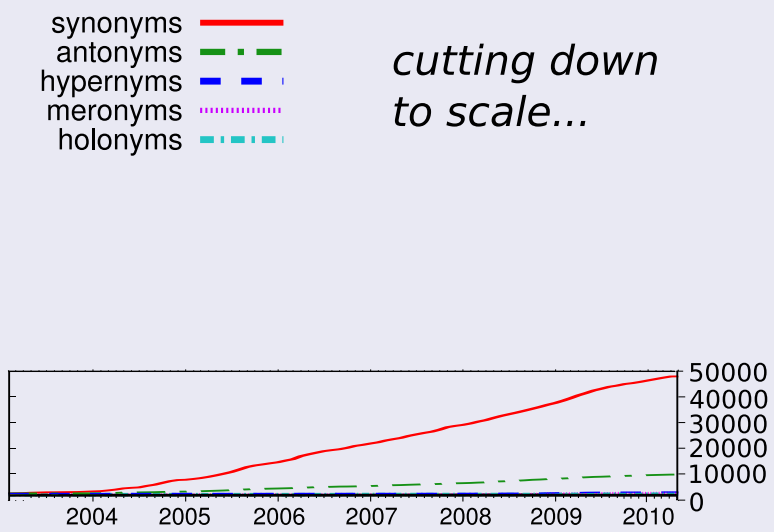
Wiktionary anglais : évolution de 2003 à 2010



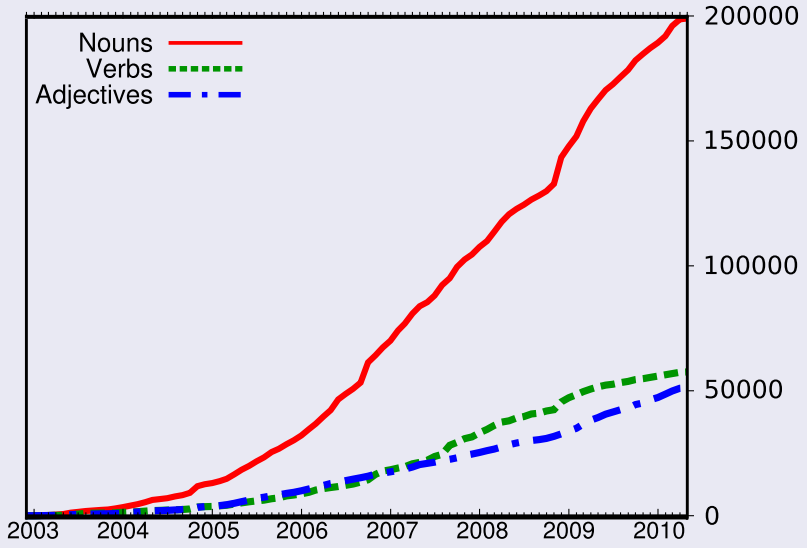
Croissance vocables vs. relations sémantiques

Wiktionary anglais : évolution de 2003 à 2010

Semantic relations (all POS taken together)



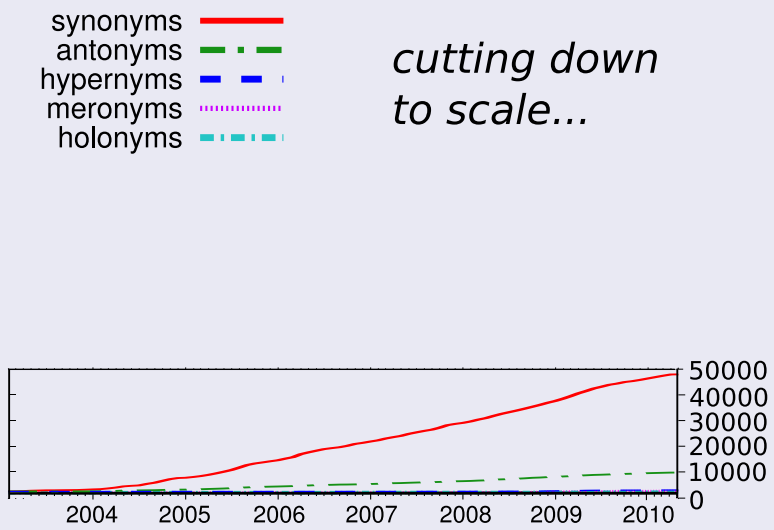
Lexemes



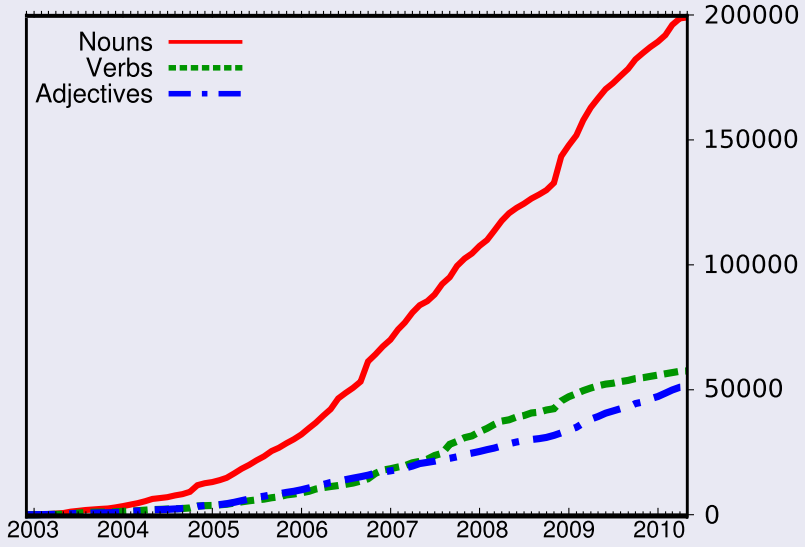
Croissance vocables vs. relations sémantiques

Wiktionary anglais : évolution de 2003 à 2010

Semantic relations (all POS taken together)



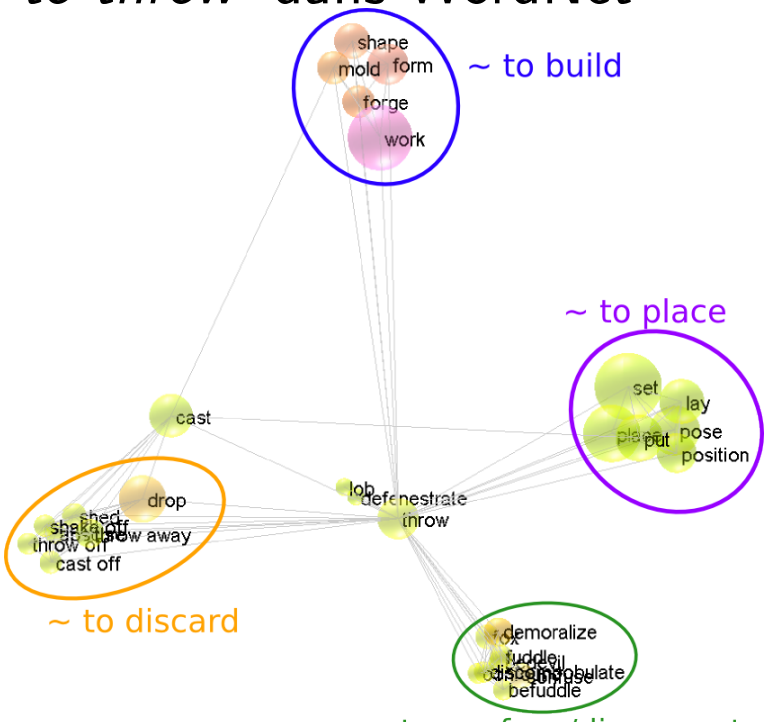
Lexemes



Proposition : aider à accélerer la densification du réseau de synonymie

2009, premières expériences (cf. [Navarro et al., 2009])

'to throw' dans WordNet

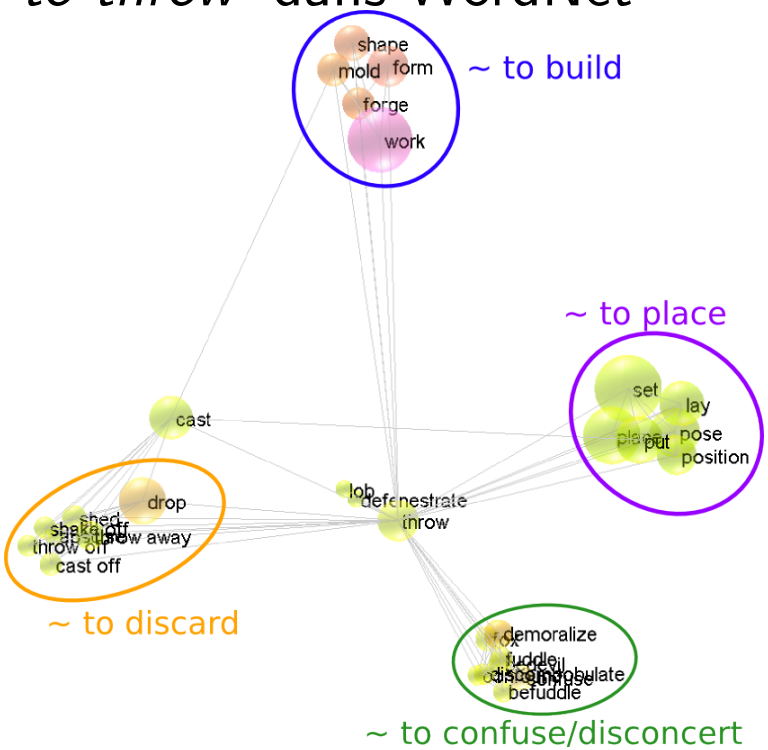


Constat et hypothèse

- Les réseaux lexicaux (e.g. réseaux de synonymie) sont des petits mondes, les sens proches forment des clusters (cf. présentation UE TAL précédente).
- Dans un réseau partiel/déficient (e.g. Wiktionary, en cours de construction), les liens manquants se trouvent (devraient se trouver) dans les clusters.

2009, premières expériences (cf. [Navarro et al., 2009])

'to throw' dans WordNet



Constat et hypothèse

- Les réseaux lexicaux (e.g. réseaux de synonymie) sont des petits mondes, les sens proches forment des clusters (cf. présentation UE TAL précédente).
- Dans un réseau partiel/déficient (e.g. Wiktionary, en cours de construction), les liens manquants se trouvent (devraient se trouver) dans les clusters.

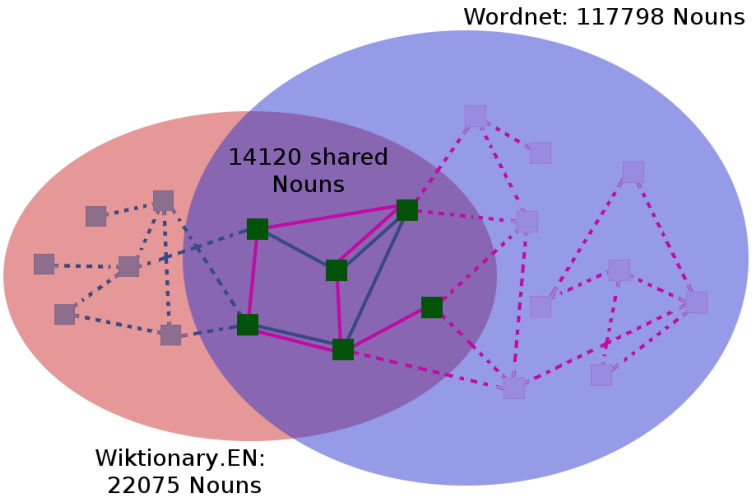
→ utilisons la proxémie pour trouver les synonymes potentiellement manquant : marches aléatoires sur les réseaux de synonymie de Wiktionary et proposition des paires de mots ≪ *les plus Prox* ≫

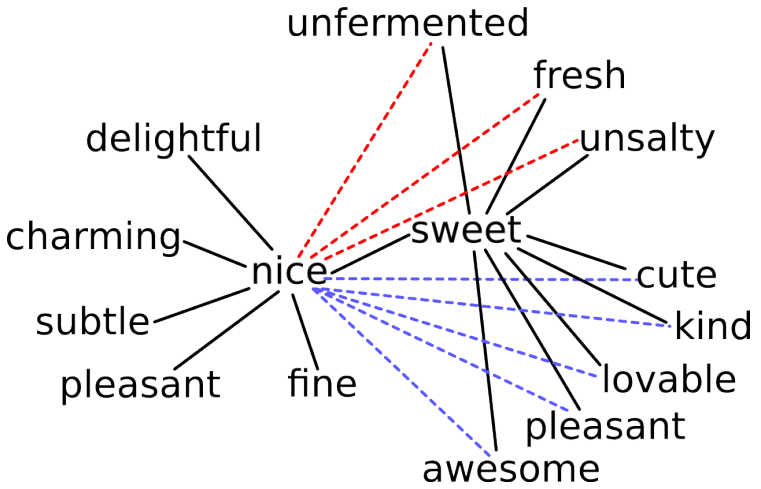
2009, premières expériences (suite)

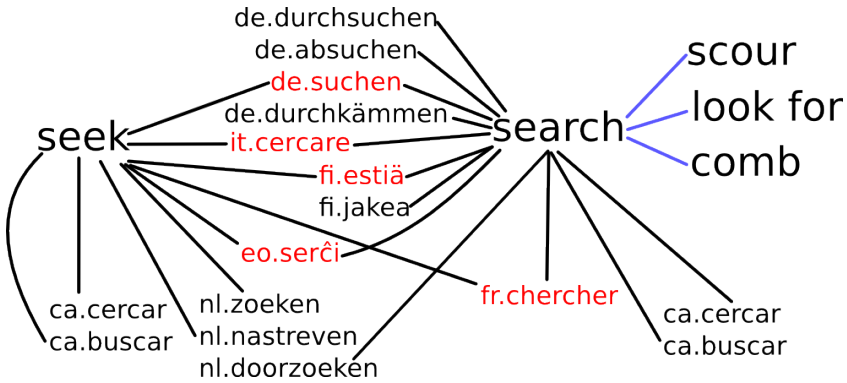
Plusieurs méthodes

- les voisins de mes voisins : transitivité
- clusters : Prox
- traductions : Jaccard

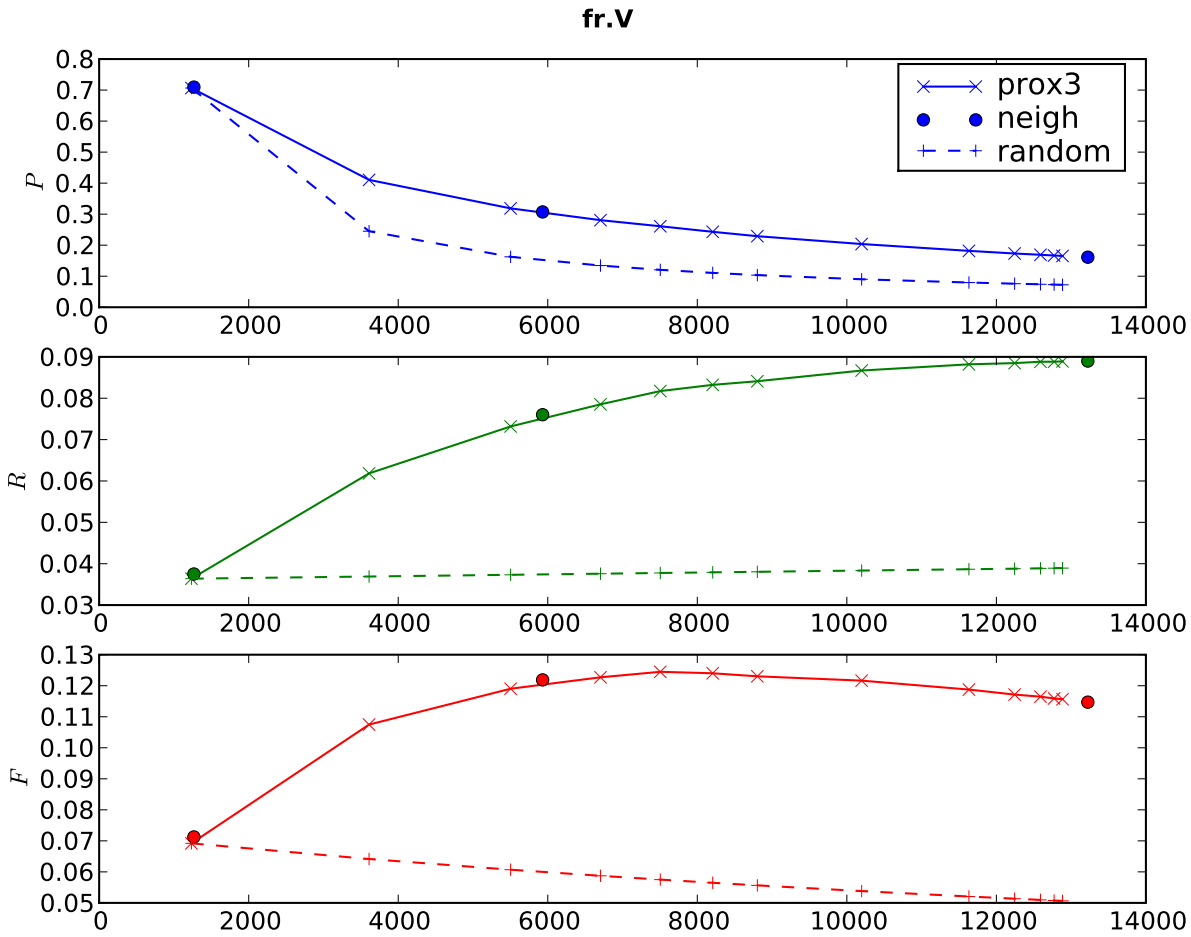
Évaluation par rapport à WordNet et DicoSyn



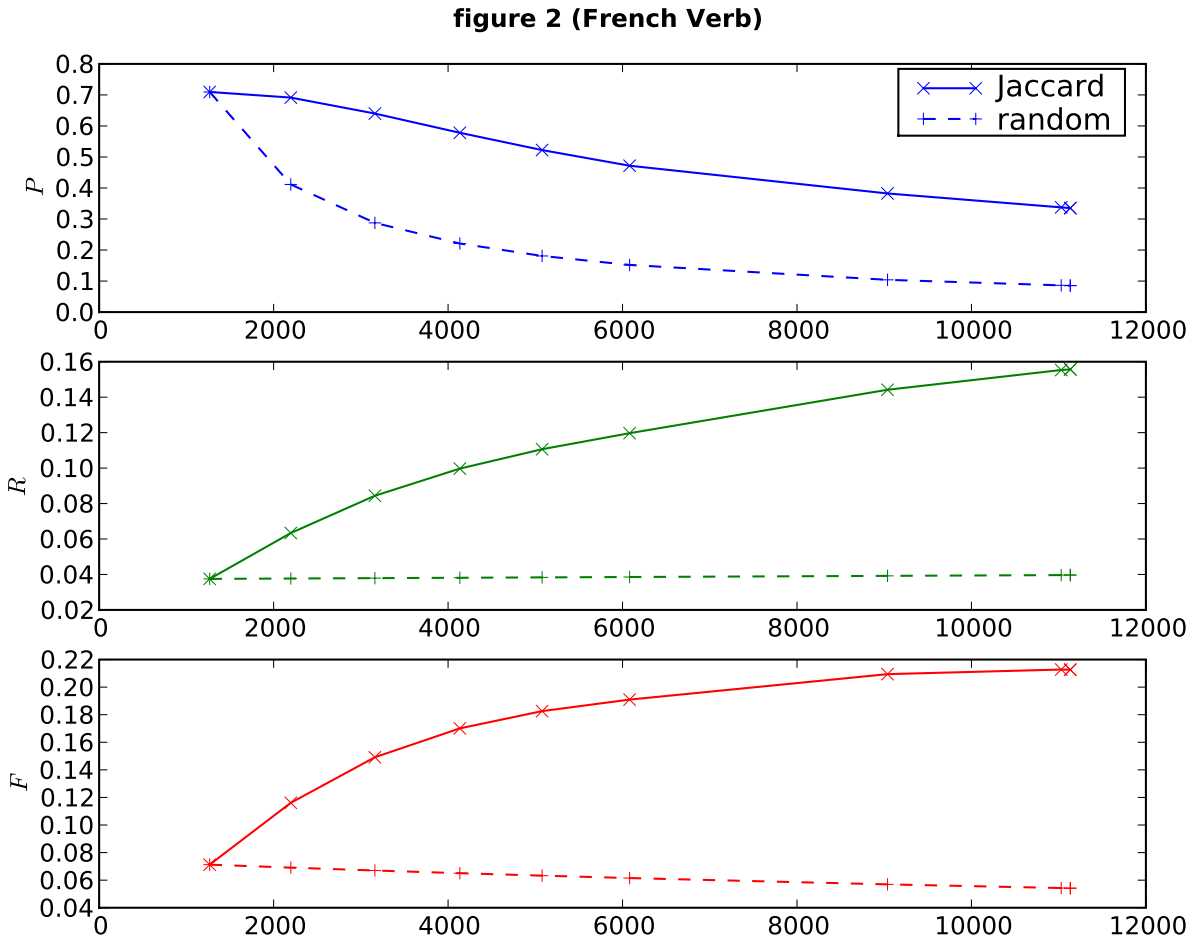




2009, résultats : Voisins et Prox



2009, résultats : Traductions

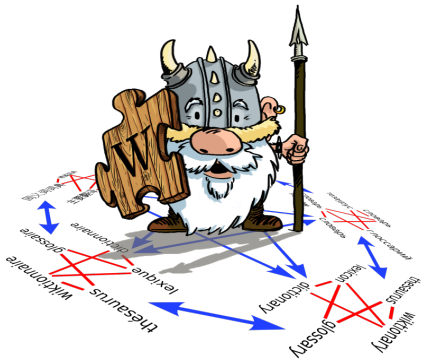


Ajout des 1000 1^{ers} liens (+55%) $\rightarrow$ seulement -2% de précision

Évaluation par rapport à des étalons

- que dire de *'to absolve'* $\leftrightarrow$ *'to forgive'* proposé par Prox, invalidé par WN ?
- que dire de *'to reduce'* $\leftrightarrow$ *'to decrease'* ou *'to cook'* $\leftrightarrow$ *'to microweave'* initialement dans Wiktionary, absent de WN ?
- meilleurs résultats sur le français que sur l'anglais : densités initiales Wiktionary/Wiktionnaire inégales ou différence de nature, densité et granularité des étalons ?

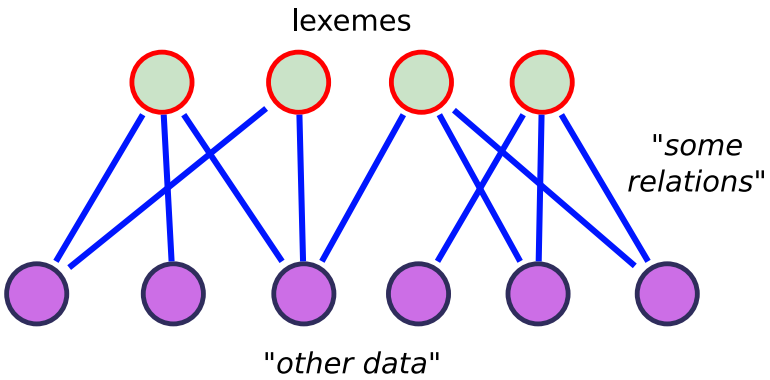
Piggybacking... sur les épaules d'un géant



Synonymes candidats bruités, problème de validation ?
Wiktionary : des contributeurs actifs...
Proposons des candidats aux contributeurs pour qu'ils les valident/invalient

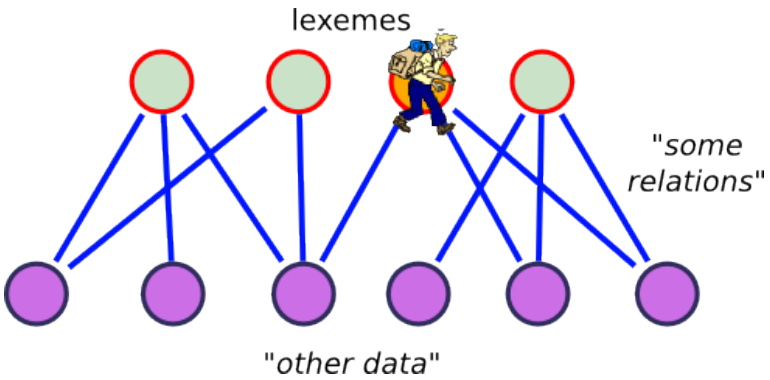
Marches aléatoires : graphes bipartis

Différentes sources de données



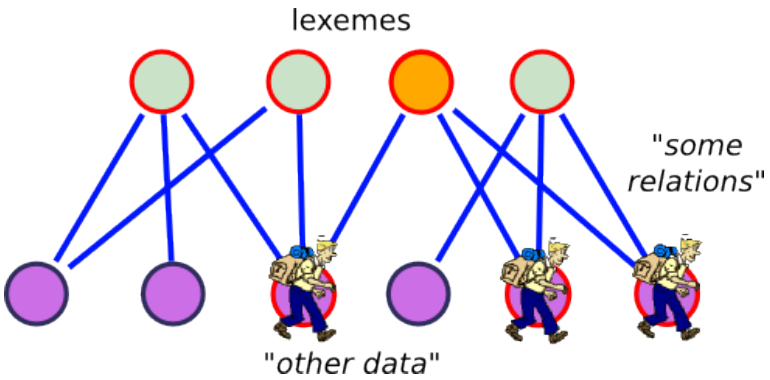
Marches aléatoires : graphes bipartis

Différentes sources de données



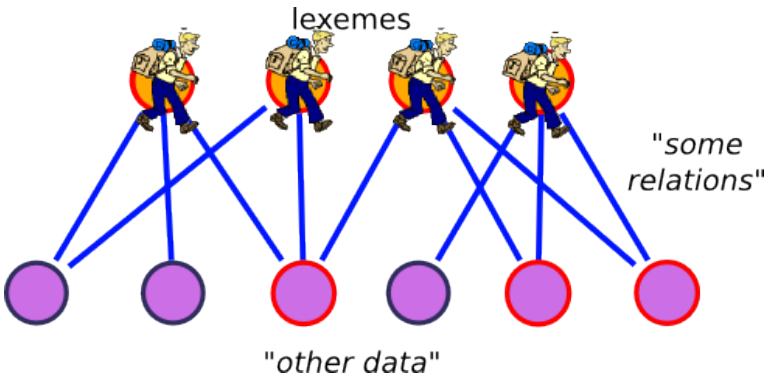
Marches aléatoires : graphes bipartis

Différentes sources de données



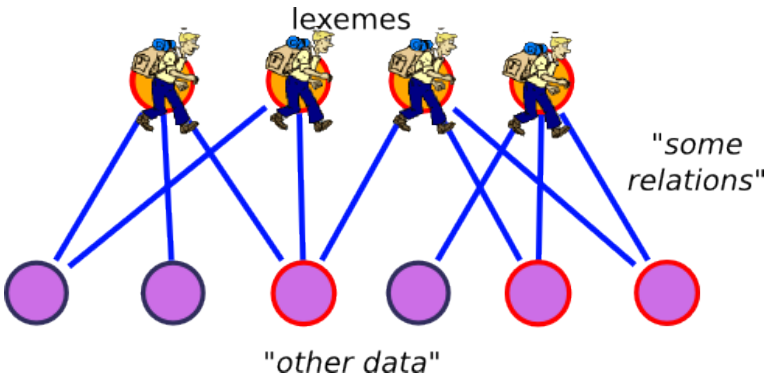
Marches aléatoires : graphes bipartis

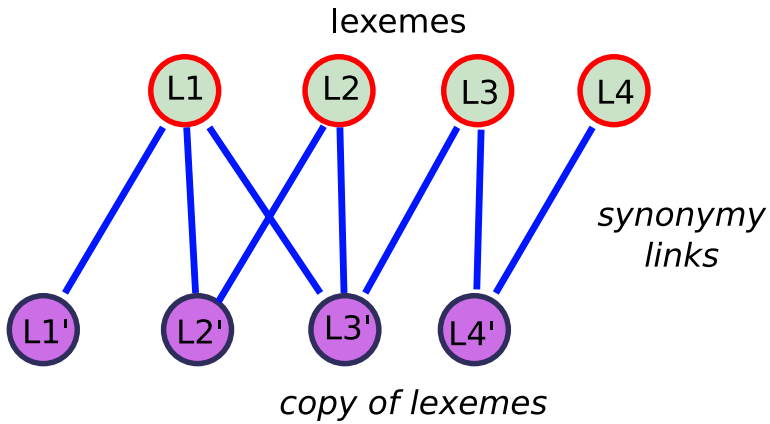
Différentes sources de données



Marches aléatoires : graphes bipartis

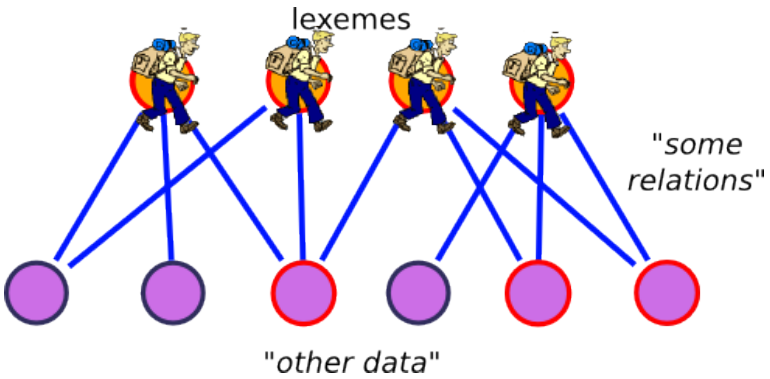
Différentes sources de données

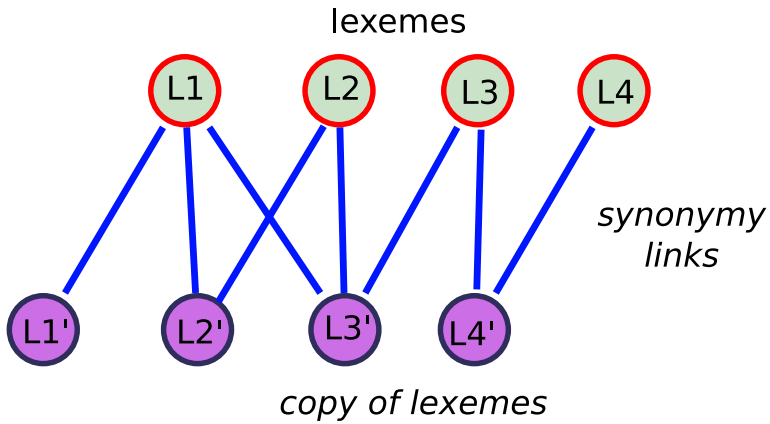


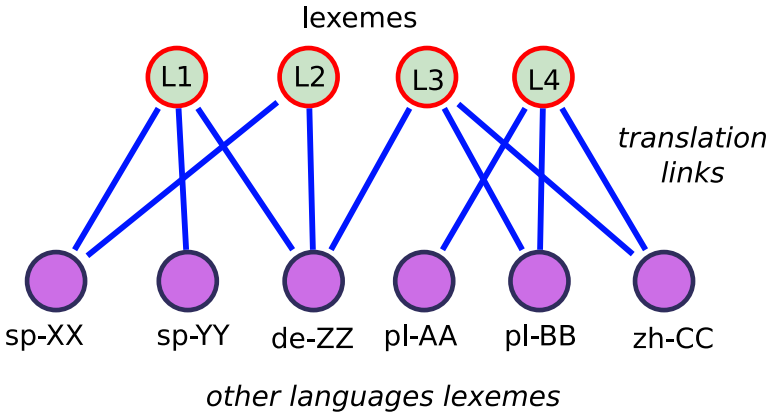


Marches aléatoires : graphes bipartis

Différentes sources de données

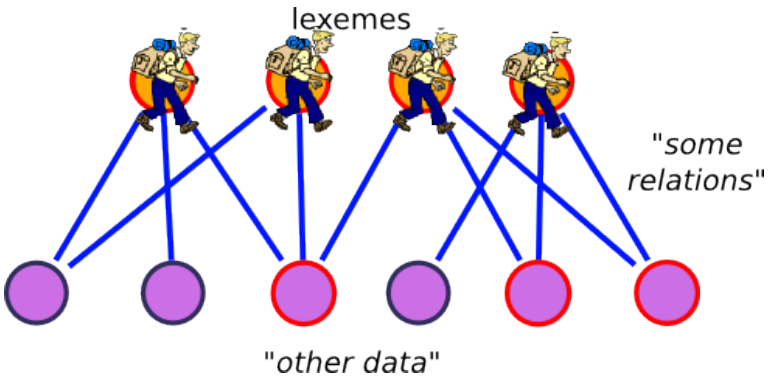


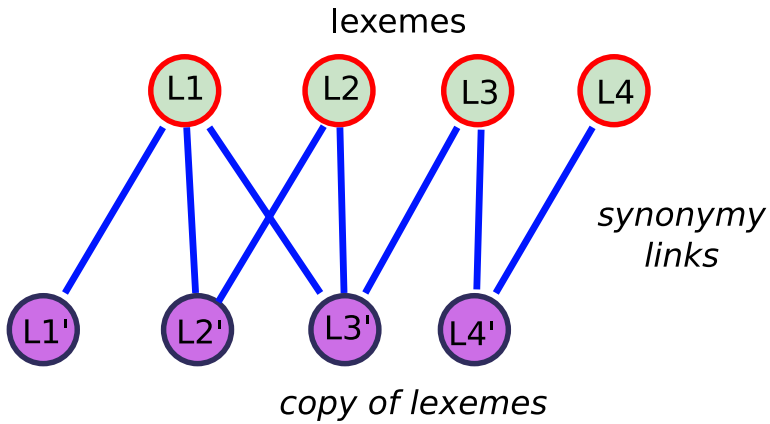


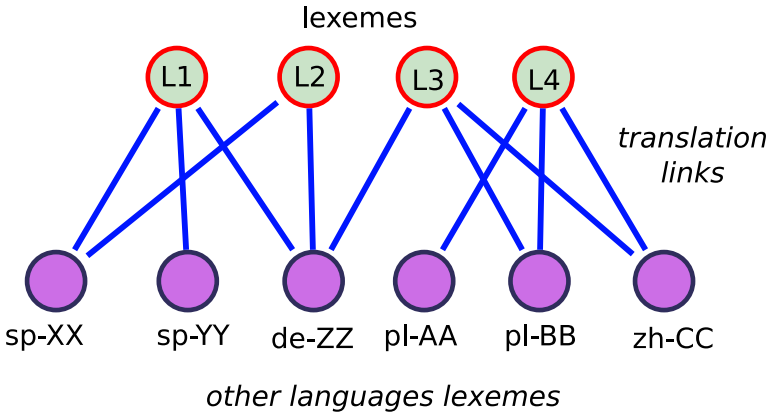


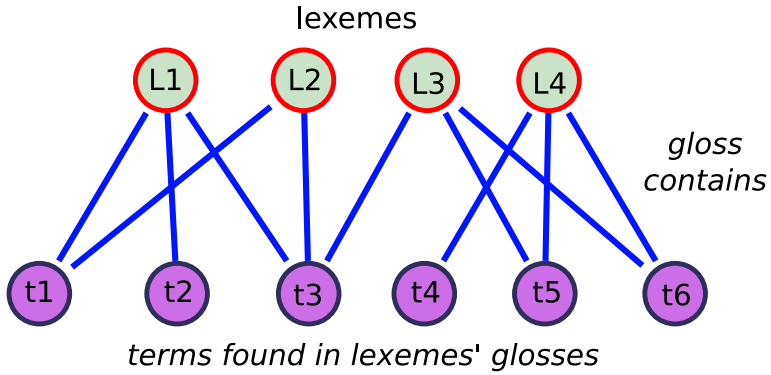
Marches aléatoires : graphes bipartis

Différentes sources de données



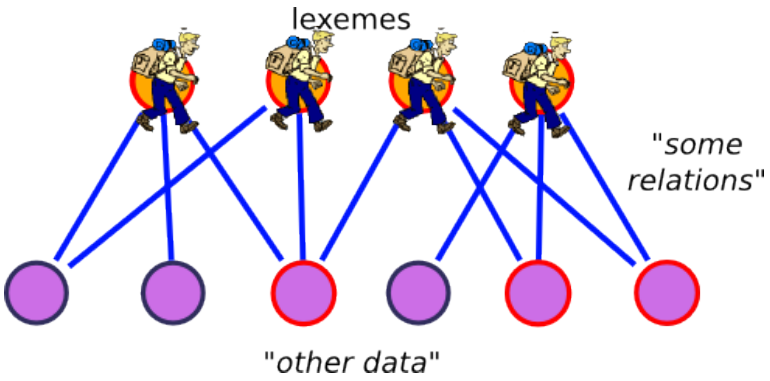


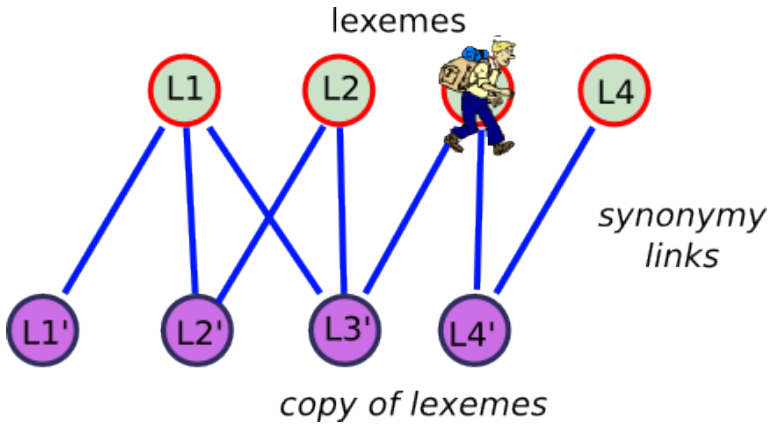


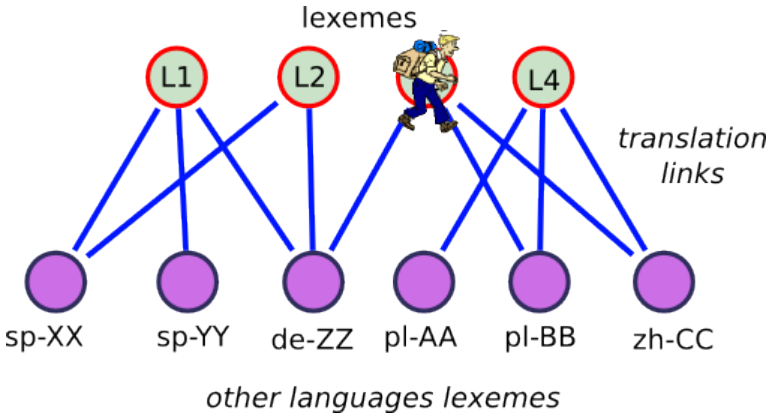


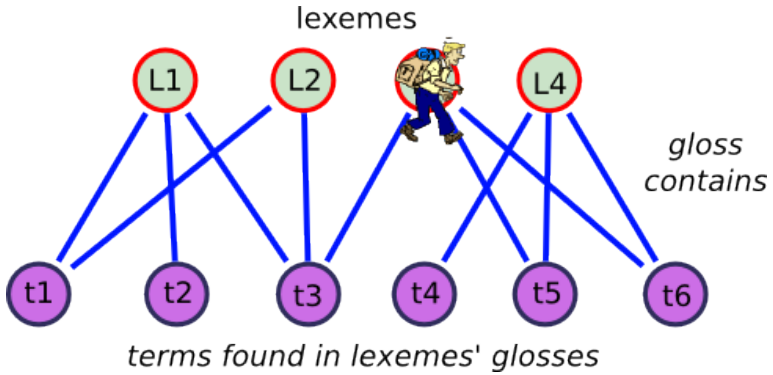
Marches aléatoires : graphes bipartis

Différentes sources de données



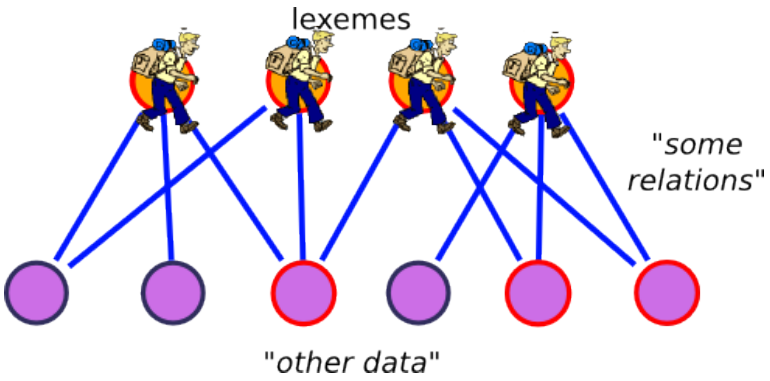


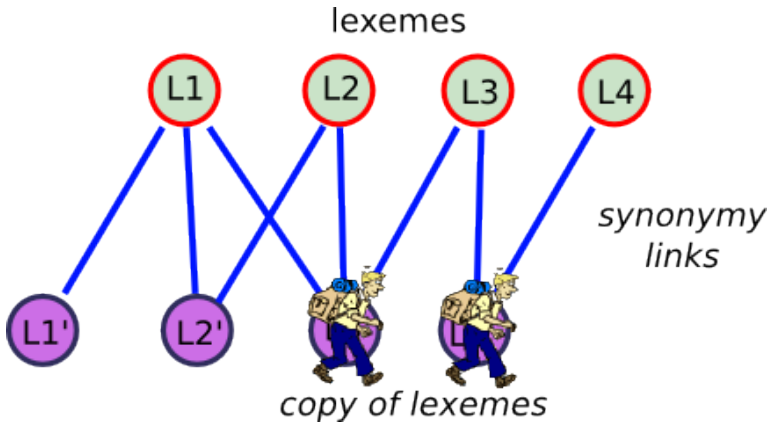


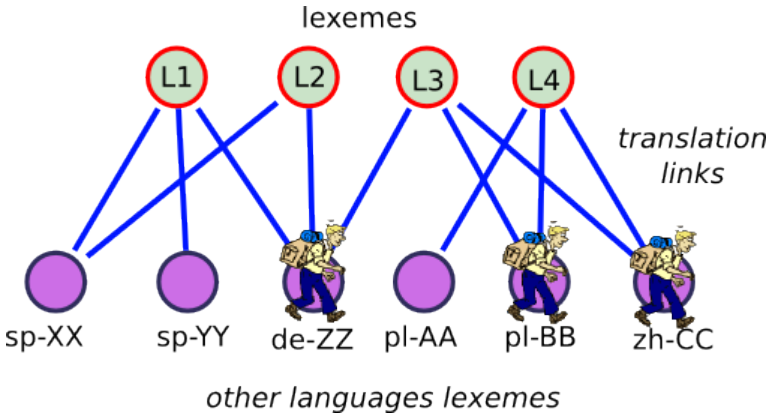


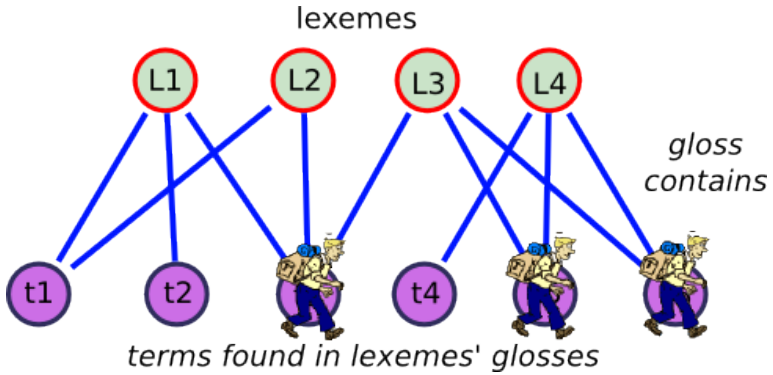
Marches aléatoires : graphes bipartis

Différentes sources de données



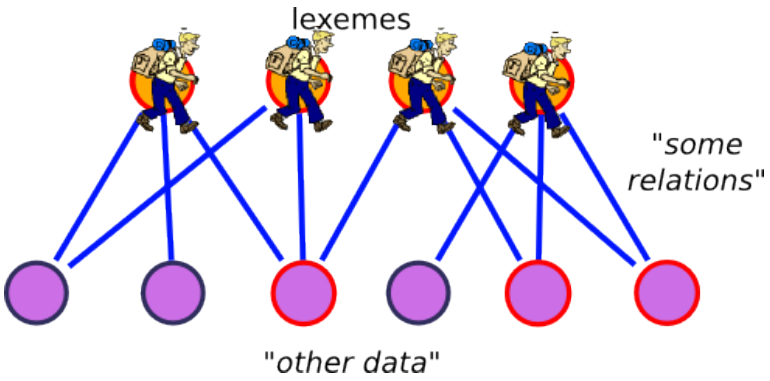


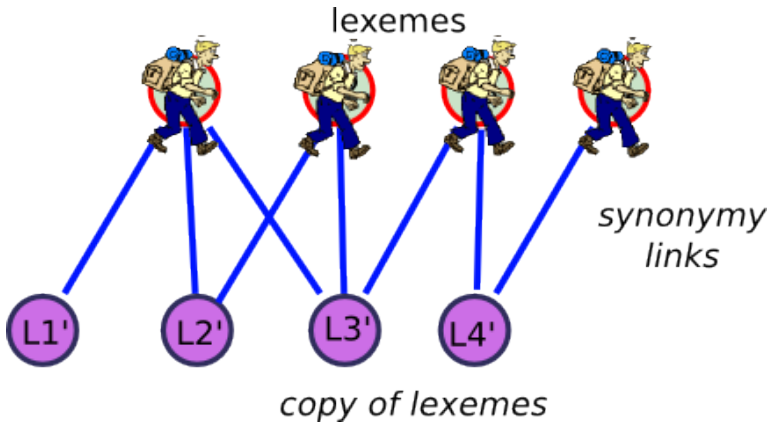


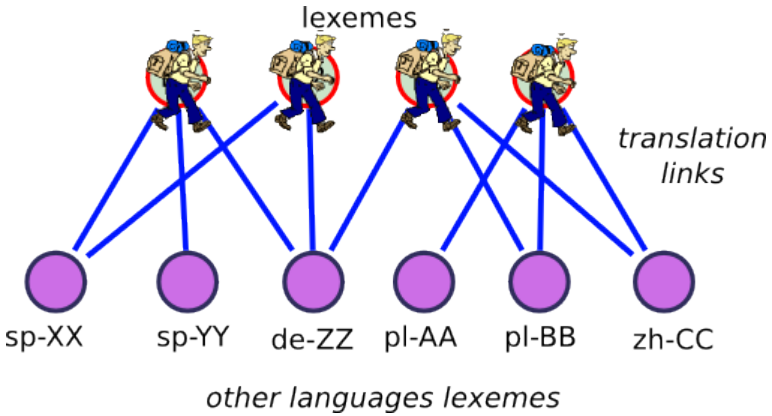


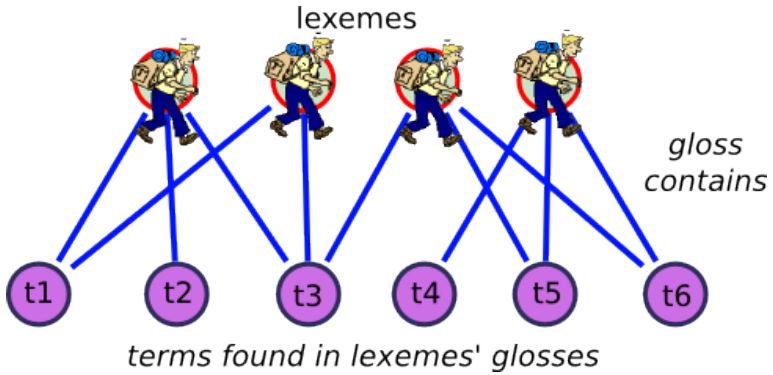
Marches aléatoires : graphes bipartis

Différentes sources de données

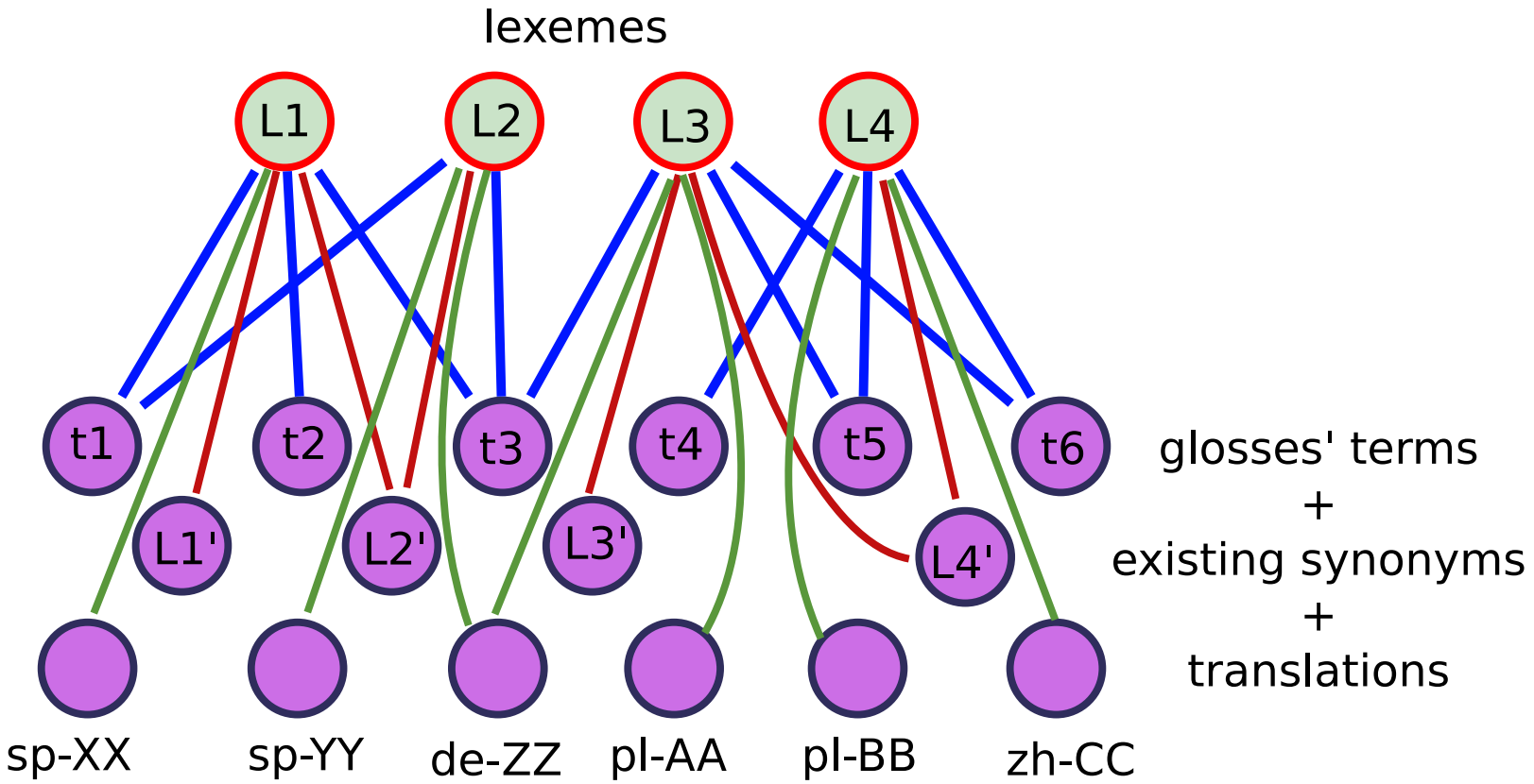








Agrégation des différentes sources de données



Pertinence des candidats proposés

Comment l'évaluer ?

- Dans une perspective semi-automatique, une liste (courte, ici : 5) de candidats est consdérée comme *acceptable* si elle contient au moins un candidat pertinent. On compte aussi celles qui en contiennent 2, 3...
- Évaluation (nous y revoila...) : compter pour combien de vocables notre méthode produit une liste acceptable.
- Candidat pertinent ? Par rapport aux golds !
- Évaluation : au moins utile pour choisir la meilleure combinaison des sources de données à utiliser

Résultats : impact des sources de données

		R_5	P_5	$\lvert N_5 \rvert$	$\lvert N_5^{+1} \rvert$	$\lvert N_5^{+2} \rvert$	$\lvert N_5^{+3} \rvert$	$\lvert N_5^{+4} \rvert$	$\lvert N_5^{+5} \rvert$
FR-V	syns	10.0	68.0	412	280	172	86	30	5
	trads	19.0	**90.4**	785	710	544	**352**	**146**	**38**
	gloses non pondérées	**95.6**	41.2	**3947**	1628	530	149	38	3
	gloses pondérées	**95.6**	44.9	**3947**	**1773**	**638**	198	45	8
	contextes syntaxiques	81.8	35.3	3378	1192	426	126	28	3
	$s + t$	25.7	**85.6**	1062	909	669	418	165	48
	$10.s + 10.t + g$	96.6	55.9	3989	2229	1161	580	216	**58**
	$10^2.s + 10^2.t + g$	96.6	55.8	3989	2226	1160	580	214	**58**
	$10^2.s + 10^2.t + 10.g + c$	**98.1**	53.2	**4053**	2158	1004	433	146	43
	$\mathbf{10^3.s + 10^3.t + 10^2.g + c}$	**98.1**	58.4	**4053**	**2368**	**1243**	**604**	**223**	53

(résultats complets dans [Sajous et al., 2011])

Résultats : impact des sources de données

		R_5	P_5	$\|N_5\|$	$\|N_5^{+1}\|$	$\|N_5^{+2}\|$	$\|N_5^{+3}\|$	$\|N_5^{+4}\|$	$\|N_5^{+5}\|$
FR-V	syns	10.0	68.0	412	280	172	86	30	5
	trads	19.0	**90.4**	785	710	544	**352**	**146**	**38**
	gloses non pondérées	**95.6**	41.2	**3947**	1628	530	149	38	3
	gloses pondérées	**95.6**	44.9	**3947**	**1773**	**638**	198	45	8
	contextes syntaxiques	81.8	35.3	3378	1192	426	126	28	3
	$s + t$	25.7	**85.6**	1062	909	669	418	165	48
	$10.s + 10.t + g$	96.6	55.9	3989	2229	1161	580	216	**58**
	$10^2.s + 10^2.t + g$	96.6	55.8	3989	2226	1160	580	214	**58**
	$10^2.s + 10^2.t + 10.g + c$	**98.1**	53.2	**4053**	2158	1004	433	146	43
	$\mathbf{10^3.s + 10^3.t + 10^2.g + c}$	**98.1**	58.4	**4053**	**2368**	**1243**	**604**	**223**	53

(résultats complets dans [Sajous et al., 2011])

	dans GS	**Propositions (noms)**
EN	Oui	*<imprisonment : captivity>, <harmony : peace>, <filth : dirt>, <antipasto : starter>, <load : burden>, <possessive : genitive>*
	Non	*<rebirth : renewal>, <fool : idiot, dummy>, <cheating : fraud>, <bypass : circumvention>, <dissimilarity : variance>, <pro : benefit>*
FR	Oui	*<ouvrage : travail>, <renom : gloire>, <emploi : fonction>, <drapeau : pavillon>, <rythme : cadence>, <roulotte : caravane>, <chinois : tamis>*
	Non	*<drogue : psychotrope>, <fantassin : bidasse>, <force : poigne>, <salade : bobard>, <W.C. : chiotte>, <us : tradition>, <bisque : soupe>*

Résultats : impact des sources de données

		R_5	P_5	$\lvert N_5 \rvert$	$\lvert N_5^{+1} \rvert$	$\lvert N_5^{+2} \rvert$	$\lvert N_5^{+3} \rvert$	$\lvert N_5^{+4} \rvert$	$\lvert N_5^{+5} \rvert$
FR-V	syns	10.0	68.0	412	280	172	86	30	5
	trads	19.0	**90.4**	785	710	544	**352**	**146**	**38**
	gloses non pondérées	**95.6**	41.2	**3947**	1628	530	149	38	3
	gloses pondérées	**95.6**	44.9	**3947**	**1773**	**638**	198	45	8
	contextes syntaxiques	81.8	35.3	3378	1192	426	126	28	3
	$s + t$	25.7	**85.6**	1062	909	669	418	165	48
	$10.s + 10.t + g$	96.6	55.9	3989	2229	1161	580	216	**58**
	$10^2.s + 10^2.t + g$	96.6	55.8	3989	2226	1160	580	214	**58**
	$10^2.s + 10^2.t + 10.g + c$	**98.1**	53.2	**4053**	2158	1004	433	146	43
	$\mathbf{10^3.s + 10^3.t + 10^2.g + c}$	**98.1**	58.4	**4053**	**2368**	**1243**	**604**	**223**	53

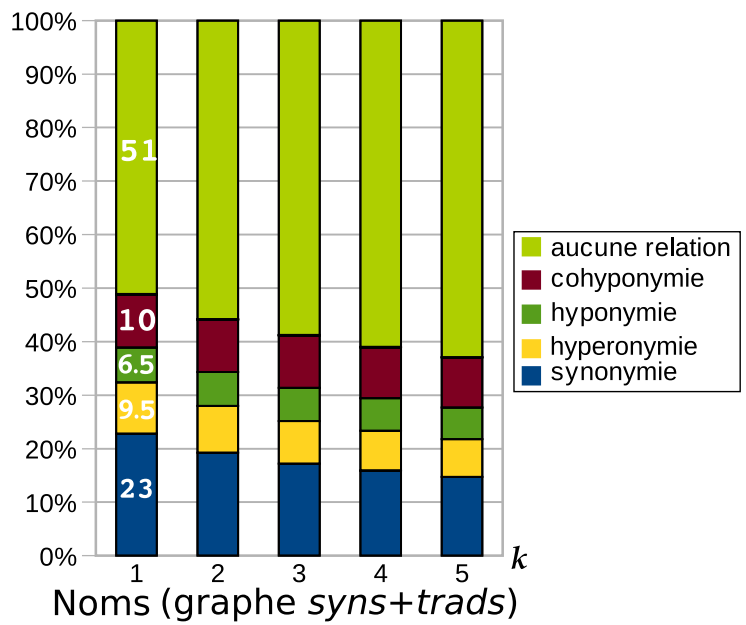
(résultats complets dans [Sajous et al., 2011])

	dans GS	**Propositions (noms)**
EN	Oui	*<imprisonment : captivity>*, *<harmony : peace>*, *<filth : dirt>*, *<antipasto : starter>*, *<load : burden>*, *<possessive : genitive>*
	Non	*<rebirth : renewal>*, *<fool : idiot, dummy>*, *<cheating : fraud>*, *<bypass : circumvention>*, *<dissimilarity : variance>*, *<pro : benefit>*
FR	Oui	*<ouvrage : travail>*, *<renom : gloire>*, *<emploi : fonction>*, *<drapeau : pavillon>*, *<rythme : cadence>*, *<roulotte : caravane>*, *<chinois : tamis>*
	Non	*<drogue : psychotrope>*, *<fantassin : bidasse>*, *<force : poigne>*, *<salade : bobard>*, *<W.C. : chiotte>*, *<us : tradition>*, *<bisque : soupe>*

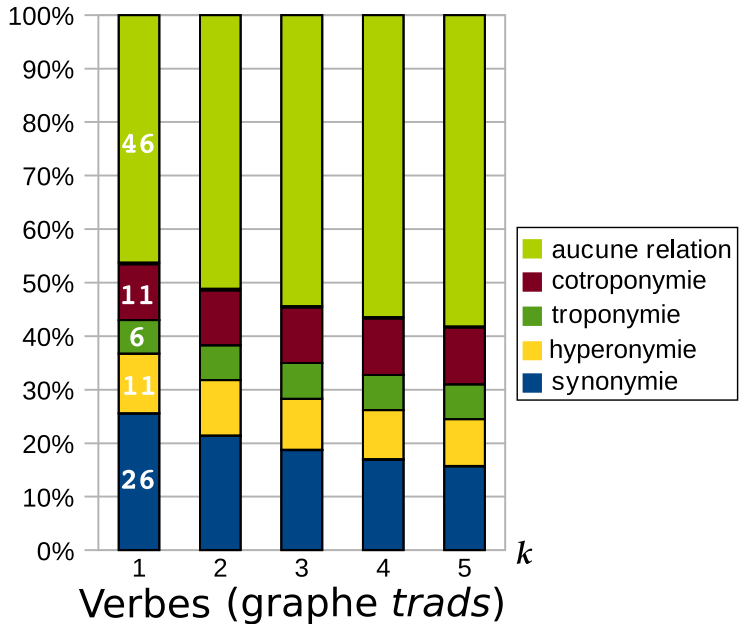
Étalons sévères !
À l'inverse, *<rongeur : mulot>* et *<sens : toucher>* validés par GS, pourtant bon candidats à l'hyperonymie (encore une fois, nature des GS : phénomène certainement moins fréquent pour WiktEN évalué par rapport à PWN !).

Caractérisation des relations capturées par marches aléatoires (voir aussi [Heylen et al., 2008])

Caractérisation des relations capturées par marches aléatoires (voir aussi [Heylen et al., 2008])



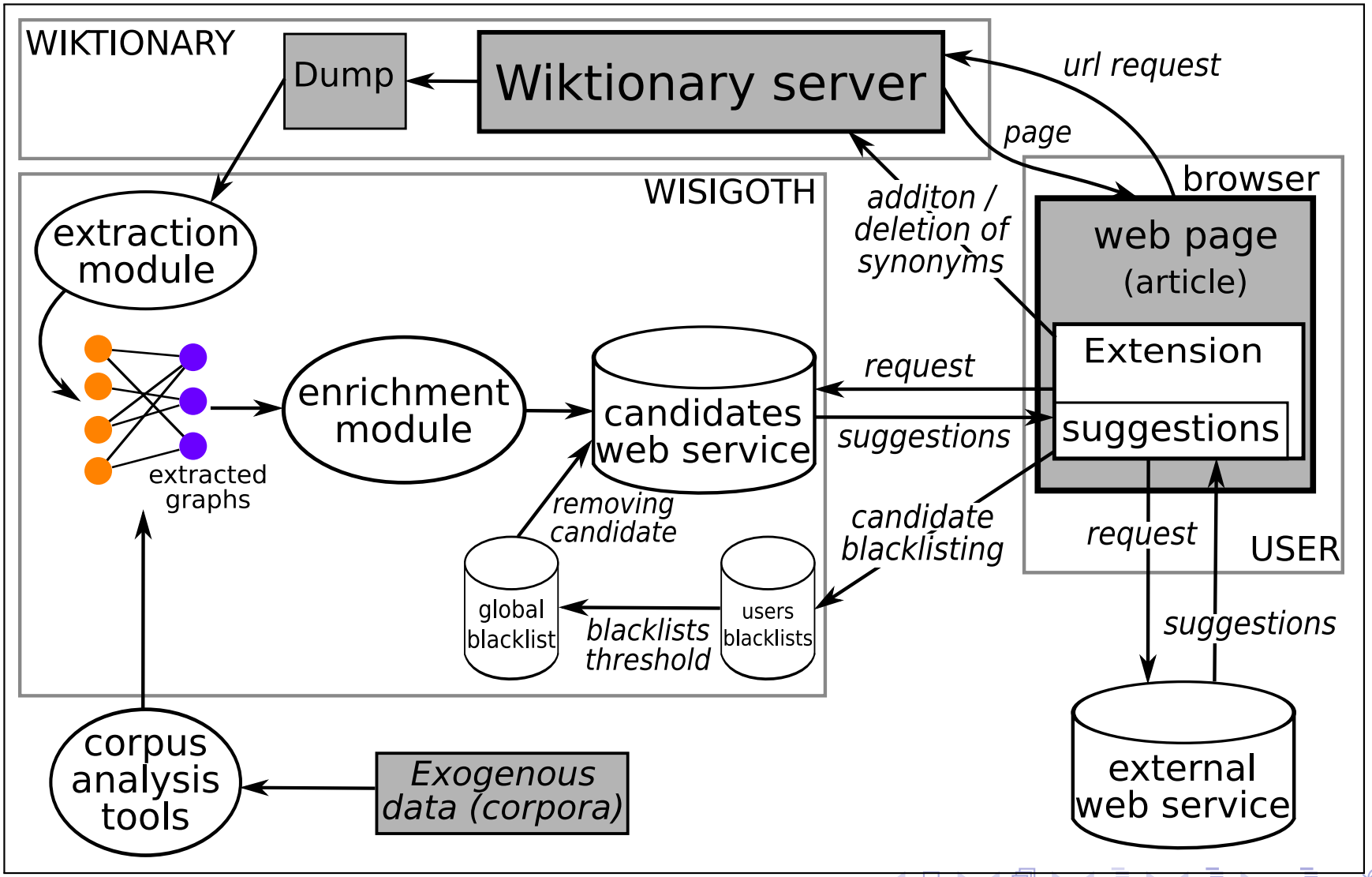
Noms (graphe *syns+trads*)



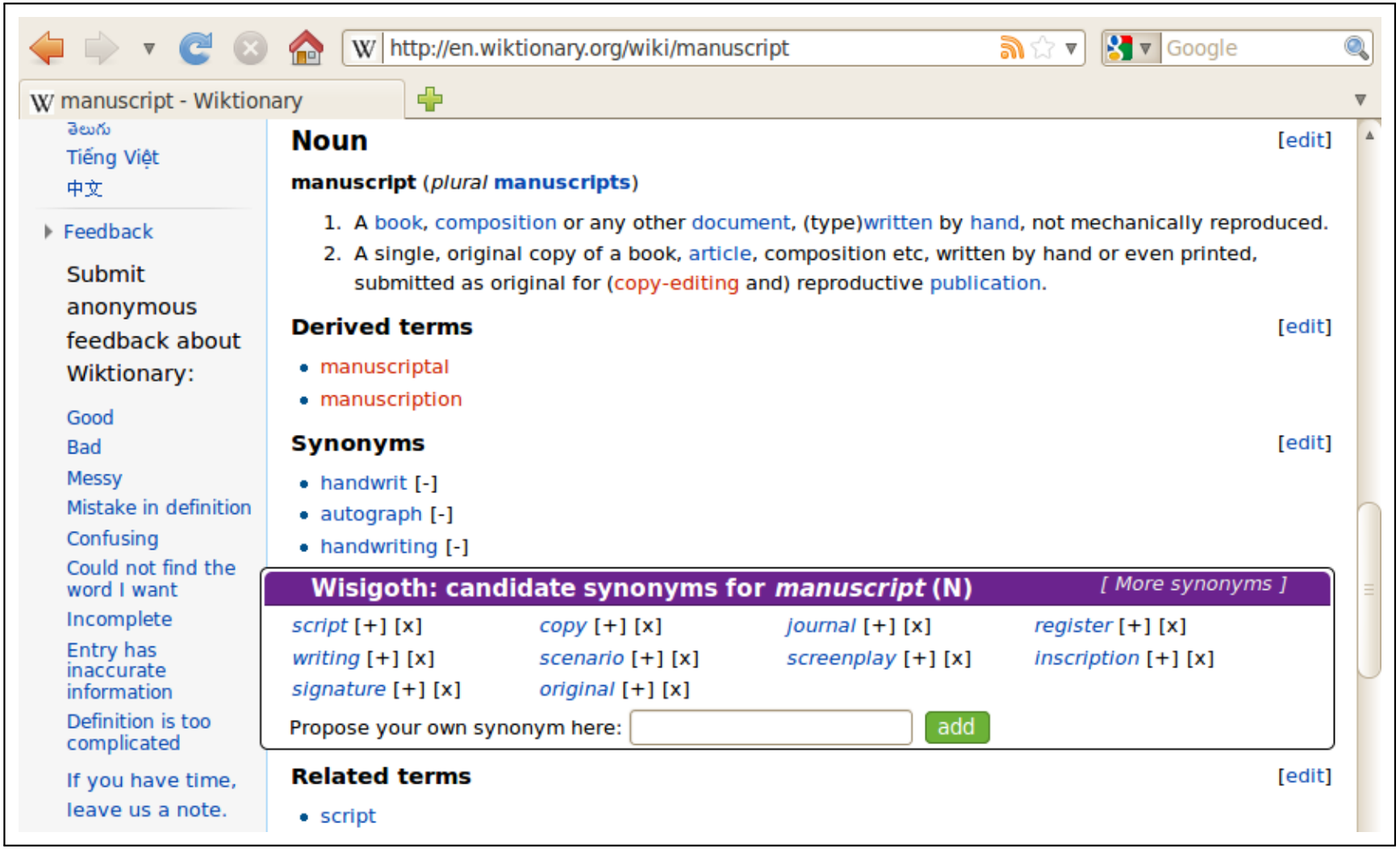
Verbes (graphe *trads*)

POS	Vocables	Candidats	Relations
N	*hound*	*greyhound*	hyponymie
	law	*rule*	hyponymie
	fool	*idiot, dummy*	cohyponymie (synset hyperonyme : {*simpleton, simple*})
V	*to represent*	*to depict*	troponymie
	to negociate	*to bargain*	troponymie
	to blame	*to incriminate*	cotroponymie (synset hyperonyme : {*to charge, to accuse*})

Architecture



Extension


http://en.wiktionary.org/wiki/manuscript
Google
manuscript - Wiktionary
తెలుగు
Tiếng Việt
中文
Feedback
Submit anonymous feedback about Wiktionary:
Good
Bad
Messy
Mistake in definition
Confusing
Could not find the word I want
Incomplete
Entry has inaccurate information
Definition is too complicated
If you have time, leave us a note.
Noun
[edit]
manuscript (plural manuscripts)
1. A book, composition or any other document, (type)written by hand, not mechanically reproduced.
2. A single, original copy of a book, article, composition etc, written by hand or even printed, submitted as original for (copy-editing and) reproductive publication.
Derived terms
[edit]
manuscriptal
manuscription
Synonyms
[edit]
handwrit [-]
autograph [-]
handwriting [-]
Wisigoth: candidate synonyms for manuscript (N)
[More synonyms]
script [+] [x]
copy [+] [x]
journal [+] [x]
register [+] [x]
writing [+] [x]
scenario [+] [x]
screenplay [+] [x]
inscription [+] [x]
signature [+] [x]
original [+] [x]
Propose your own synonym here:
add
Related terms
[edit]
script

Ressources lexicales : travaux connexes

Tendances actuelles

- Alignement de ressources complémentaires (e.g. WordNet et Framenet [Baker and Fellbaum, 2009])
- Construire automatiquement des ressources volumineuses et ne pas (ou peu) les évaluer

Ressources récentes disponibles

- OntoWiktionary, an ontology from the collaborative online dictionary Wiktionary [Meyer and Gurevych, 2012] pour l'anglais, l'allemand et le russe
- UBY, A Large-Scale Unified Lexical-Semantic Resource [Gurevych et al., 2012] : mapping de 8 ressources (Wiktionary, Wikipedia, OmegaWiki, WordNet, VerbNet, FrameNet, GermaNet) pour l'anglais et l'allemand (alignement au niveau des sens)
- PanDictionary [Mausam et al., 2009] : graphe de traductions ancrées au niveau des sens (après désamb.)
- BabelNet [Navigli and Ponzetto, 2010] : WN + Wikipédia + trad. automatique $\rightarrow$ "dictionnaire encyclopédique"
- Toujours rien pour le français

(No) ? future work

WISIGOTH soumis aux aléas de :

- version de Firefox
- structure du code wiki
- layout HTML

Future work

→ mise à disposition d'une version XML actualisée du Wiktionnaire (et du parseur)

Future work

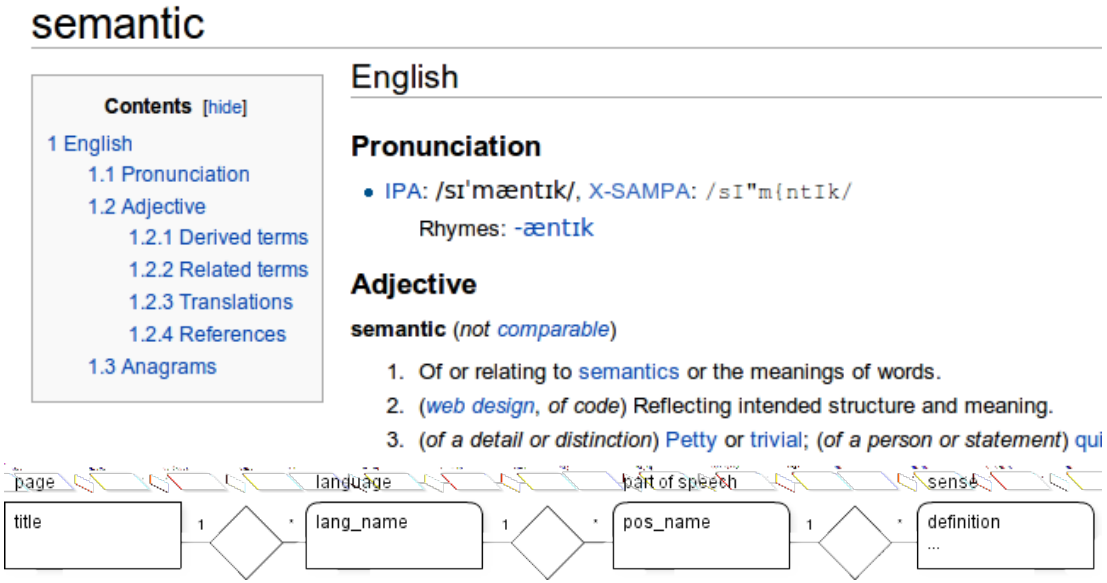
→ mise à disposition d'une version XML actualisée du Wiktionnaire (et du parseur)

[Hellmann et al., 2012] : *"This is done by so called extraction templates(ET) (not to be confused with the templates of wikitext). Each possible section in the Wiktionary page layout (i.e. each linguistic property) has an ET configured. The idea is to provide a declarative and intuitive way to encode what to extract."*

Future work

→ mise à disposition d'une version XML actualisée du Wiktionnaire (et du parseur)

[Hellmann et al., 2012] : *"This is done by so called extraction templates(ET) (not to be confused with the templates of wikitext). Each possible section in the Wiktionary page layout (i.e. each linguistic property) has an ET configured. The idea is to provide a declarative and intuitive way to encode what to extract."*




```
=== Synonyms ===
* [[ building ]]
* [[ company ]]
```

```
=== Synonyms ===
(* [[\ $target ]]
 )+
```

Future work (2)

- traductions ?
- production d'un lexique morphophonologique (vers un BDlex libre ?)

Français [modifier]

Étymologie

Du grec ancien λεξικόν, *lexikon* (« livre de mots »).

Nom commun

lexique /lɛk.sik/ *masculin*

1. Répertoire d'une catégorie de mots sauf ceux grammaticaux tels les prépositions et les conjonctions
2. Dictionnaire des locutions et formes propres à certains auteurs.
   - *Le **lexique** de Platon.*

Singulier	Pluriel
lexique	lexiques
/lɛk.sik/	

References I

Baker, C. F. and Fellbaum, C. (2009).
Wordnet and framenet as complementary resources for annotation.
In *Proceedings of the Third Linguistic Annotation Workshop*, pages 125–129, Suntec, Singapore. Association for Computational Linguistics.

Chamberlain, J., Kruschwitz, U., and Poesio, M. (2009).
Constructing an anaphorically annotated corpus with non-experts : Assessing the quality of collaborative annotations.
In *Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, pages 57–62, Singapore.

Curran, J. R. and Moens, M. (2002).
Improvements in Automatic Thesaurus Extraction.
In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, pages 59–66.

Fellbaum, C., editor (1998).
WordNet : An Electronic Lexical Database.
MIT Press.

References II


Gao, H., Barbier, G., and Goolsby, R. (2011).
Harnessing the crowdsourcing power of social media for disaster relief.
IEEE Intelligent Systems, 26 :10–14.


Giles, J. (2005).
Internet Encyclopaedias go Head to Head.
Nature, 438 :900–901.


Gurevych, I., Matuschek, M., Nghiem, T.-D., Eckle-Kohler, J., Hartmann, S., and Meyer, C. (2012).
Navigating sense-aligned lexical-semantic resources : The web interface to UBY.
In Jancsary, J., editor, *Proceedings of KONVENS 2012*, pages 194–198. ÖGAI.
Main track : poster presentations.


Hearst, M. A. (1992).
Automatic acquisition of hyponyms from large text corpora.
In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING)*, pages 539–545, Nantes.

References III


Hellmann, S., Brekle, J., and Auer, S. (2012).
Leveraging the crowdsourcing of lexical resources for bootstrapping a linguistic data cloud.
In *JIST*.


Heylen, K., Peirsman, Y., Geeraerts, D., and Speelman, D. (2008).
Modelling Word Similarity : an Evaluation of Automatic Synonymy Extraction Algorithms.
In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.


Howe, J. (2006).
The rise of crowdsourcing.
Wired Magazin, 14.06.


Jacquin, C., Desmontils, E., and Monceaux, L. (2007).
French EuroWordNet Lexical Database Improvements.
In *CICLing '07 : Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 12–22. Springer-Verlag.

References IV


Kilgarriff, A. (1998).
Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs.
Computer Speech & Language, 12(4) :453–472.


Lafourcade, M. (2007).
Making People Play for Lexical Acquisition with the JeuxDeMots prototype.
In *SNLP'07 : 7th International Symposium on Natural Language Processing*, Pattaya, Thailand.


Mausam, Soderland, S., Etzioni, O., Weld, D., Skinner, M., and Bilmes, J. (2009).
Compiling a massive, multilingual dictionary via probabilistic inference.
In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 262–270, Suntec, Singapore. Association for Computational Linguistics.

References V


Meyer, C. M. and Gurevych, I. (2012).
Ontowiktionary - constructing an ontology from the collaborative online dictionary wiktionary.
In Pazienza, M. T. and Stellato, A., editors, *Semi-Automatic Ontology Development : Processes and Resources*, chapter 6, pages 131–161. IGI Global, Hershey, PA, USA.


Murray, G. C. and Green, R. (2004).
Lexical Knowledge and Human Disagreement on a WSD Task.
Computer Speech & Language, 18(3) :209–222.


Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., and Huang, C.-R. (2009).
Wiktionary and NLP : Improving synonymy networks.
In *Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, pages 19–27, Suntec, Singapore. Association for Computational Linguistics.

References VI


Navigli, R. and Ponzetto, S. P. (2010).
Babelnet : Building a very large multilingual semantic network.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.


Palmer, M., Dang, H. T., and Fellbaum, C. (2007).
Making Fine-Grained and Coarse-Grained Sense Distinctions, Both Manually and Automatically.
Natural Language Engineering, 13(2) :137–163.


Pedersen, B. S. (2010).
Releasing lexical resources as open source - pros and cons.
FLaReNet Forum 2010 : Language Resources of the future - the future of Language Resources. Barcelona, Spain.


Quinn, A. and Bederson, B. (2009).
A taxonomy of distributed human computation.
Technical report, HCIL.

References VII


Sajous, F., Navarro, E., and Gaume, B. (2011).
Enrichissement de lexiques sémantiques approvisionnés par les foules : le système WISIGOTH appliqué à Wiktionary.
TAL, 52(1) :11–35.


Sajous, F., Navarro, E., Gaume, B., Prévot, L., and Chudy, Y. (2010).
Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources : Piggybacking onto Wiktionary.
In Loftsson, H., Rögnvaldsson, E., and Helgadóttir, S., editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 332–344. Springer Berlin / Heidelberg.


Sajous, F., Navarro, E., Gaume, B., Prévot, L., and Chudy, Y. (2012).
Semi-automatic enrichment of crowdsourced synonymy networks : the WISIGOTH system applied to Wiktionary.
Language Resources and Evaluation, pages 1–34.


Sekine, S. (2010).
We Desperately Need Linguistic Resources ! –Based on the Users' Point of View.

References VIII


Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008).
Cheap and Fast—but is it good ? : Evaluating Non-Expert Annotations for Natural Language Tasks.
In *EMNLP '08 : Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Morristown, NJ, USA. Association for Computational Linguistics.


Soria, C., Monachini, M., Bertagna, F., Calzolari, N., Huang, C.-R., Hsieh, S.-K., Marchetti, A., and Tesconi, M. (2009).
Exploring Interoperability of Language Resources : the Case of Cross-Lingual Semi-Automatic Enrichment of Wordnets.
Language Resources and Evaluation, 40(1) :87–96.


Tufiș, D. and Cristea, D. (2002).
Methodological issues in building the Romanian Wordnet and consistency checks in BalkaNet.
In Christodoulakis, D. N., Kunze, C., and Lemnitzer, L., editors, *Proceedings of LREC 2002 Workshop on Wordnet Structures and Standardisation*, pages 35–41, Las Palmas, Spain.

References IX


Van der Plas, L. and Bouma, G. (2005).
Syntactic Contexts for Finding Semantically Related Words.
In van der Wouden, T., Poß, M., Reckman, H., and Cremers, C., editors, *Computational Linguistics in the Netherlands 2004 : Selected papers from the fifteenth CLIN meeting*, volume 4 of *LOT Occasional Series*. Utrecht University.


von Ahn, L. and Dabbish, L. (2004).
Labeling images with a computer game.
In *Proceedings of the CHI2004 (Conference on Human Factors in Computing Systems)*, pages 319–326.


Vossen, P., editor (1998).
EuroWordNet : a multilingual database with lexical semantic networks.
Kluwer Academic Publishers, Norwell, MA, USA.


Zesch, T. (2010).
What's the Difference ? –Comparing Expert-Built and Collaboratively-Built Lexical Semantic Resources.

References X

Zesch, T., Müller, C., and Gurevych, I. (2008).
Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary.
In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Marrakech.