Cours 3 : sémantique lexicale

Wordnet :

- une base de données lexicographiques pour l'anglais

- développée par le Cognitive Science Laboratory (Princeton University) sous la direction de George A. Miller et Christiane Fellbaum

http://www.cogsci.princeton.edu/~wn/

Wordnet : c'est quoi ?

- Wordnet : un réseau lexical qui couvre la grande majorité des noms, verbes, adjectifs et adverbes de la langue anglaise.
- Developpement (manuel) commence en 1985
- Première version de Wordnet : 1993 jusqu'à wordnet 2.1 : 3/2005
- Ressource libre et bien documentée
- Budget : $3 million du gouvernement ou des agences intéressées par la traduction

Eléments de base de Wordnet

- Wordnet définit le sens des mots par deux moyens :
  - Les synsets *(synonym set)*: ensembles de synonymes
  - Les relations entre synsets : hyperonymie- hyponymie (is-a), méronymie, implication, dérivation morphologique ....
- Exemple : {man, adult male}
  - hyperonymes :

    => {person, individual, someone, mortal, human, soul}

    => {living thing, life form, organism, being}

    => entity

    => {causal agent, cause, causal agency}

    => entity
  - hyponymes : {blackman, bachelor, dandy, boyfriend...}
  - méronyme : {adult male body}

Wordnet

- Wordnet : vaste réseau de 150 000 mots
  - organisés en 115 000 synsets
  - compressé en 12 megaoctets

- Wordnet est donc **un réseau lexical** où :
  - les synsets sont les noeuds
  - les relations entre synsets sont les arcs

Les synsets

- un synset = ensemble de mots quasi-synonymes, sorte de « classe d'équivalence » sémantique, représentant un sens (un concept) particulier
- Les mots ayant plusieurs sens appartiennent à plusieurs synsets
- Chaque synset est accompagné :
  - d'une description du sens qu'il représente : “glose”
  - d'exemples d'usage
- Exemple : synset : {building, edifice}
  - description du sens (glose) : a structure that has a roof and walls and stands more or less permanently in one place
  - exemples :
    - *there was a three-story building on the corner*
    - *it was an imposing edifice*

Traitement de la polysémie

- Les mots ayant plusieurs sens appartiennent à plusieurs synsets
- Leurs sens sont ordonnés par ordre de fréquence
- Eclatement des sens
  - 21 sens pour *cut*
    - sense 1 {cut, separate with an instrument}
    - sense 6 {cut, style, tailor}
  - 8 sens de *book*
    - sense 2 : {book, volume}
      - *a book as a physical object: a number of pages bound together*
      - *he used a large book as a doorstop"*
    - sense 8 {book}
      - *a number sheets (ticket or stamps etc.) bound together on one edge;*
      - *"he bought a book of stamps"*

Un exemple complet : *building* (nom) a 4 sens

- *building1*
  - {building, edifice}
  - a structure that has a roof and walls and stands more or less permanently in one place
  - "there was a three-story building on the corner"; "it was an imposing edifice"
- *building2*
  - {construction, building}
  - the act of constructing something
  - "during the construction we had to take a detour"; "his hobby was the building of boats"
- *building3*
  - construction, building
  - the commercial activity involved in repairing old structures or constructing new ones
  - "their main business is home construction"; "workers in the building trades"
- *building4*
  - building
  - the occupants of a building
  - "the entire building complained about the noise"

Hiérarchie de building

Sense 1
building, edifice -- (a structure that has a roof and walls)
=> structure, construction -- (a thing constructed; a complex construction or entity),
=> artifact, artefact -- (a man-made object),
=> object, inanimate object, physical object -- (a nonliving entity),
=> entity -- (something having concrete existence; living or nonliving),
Sense 2
construction, building -- (the act of constructing or building something; "his hobby was the building of boats")
=>
creating from raw materials -- (the act of creating something that is different from the materials th
=> creation -- (the human act of creating),
=> activity -- (any specific activity or pursuit; "they avoided all recreational activity"),
=> act, human action,
human activity -- (something that people do or cause to happen),
Sense 3
construction, building -- (the commerical activity involved in constructing buildings)
=> commercial enterprise, business enterprise, business -- (the activity of providing goods and services involving financial and commercial and industrial aspects; "computers are now widely used in business"),
=> commerce, commercialism, mercantilism -- (activities having the objective of supplying commodities),
=> group action -- (action taken by a group of people),
=> act, human action, human activity -- (something that people do or cause to happen),

Hiérarchie de building

Sense 4 :

building4 : gathering, assemblage (a group of persons together in one place)

=> social group (people sharing some social relation)

=> group, grouping (any number of entities (members) considered as a unit)

=> abstraction (a general concept formed by extracting common features from specific examples)

=> abstract entity (an entity that exists only abstractly)

=> entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

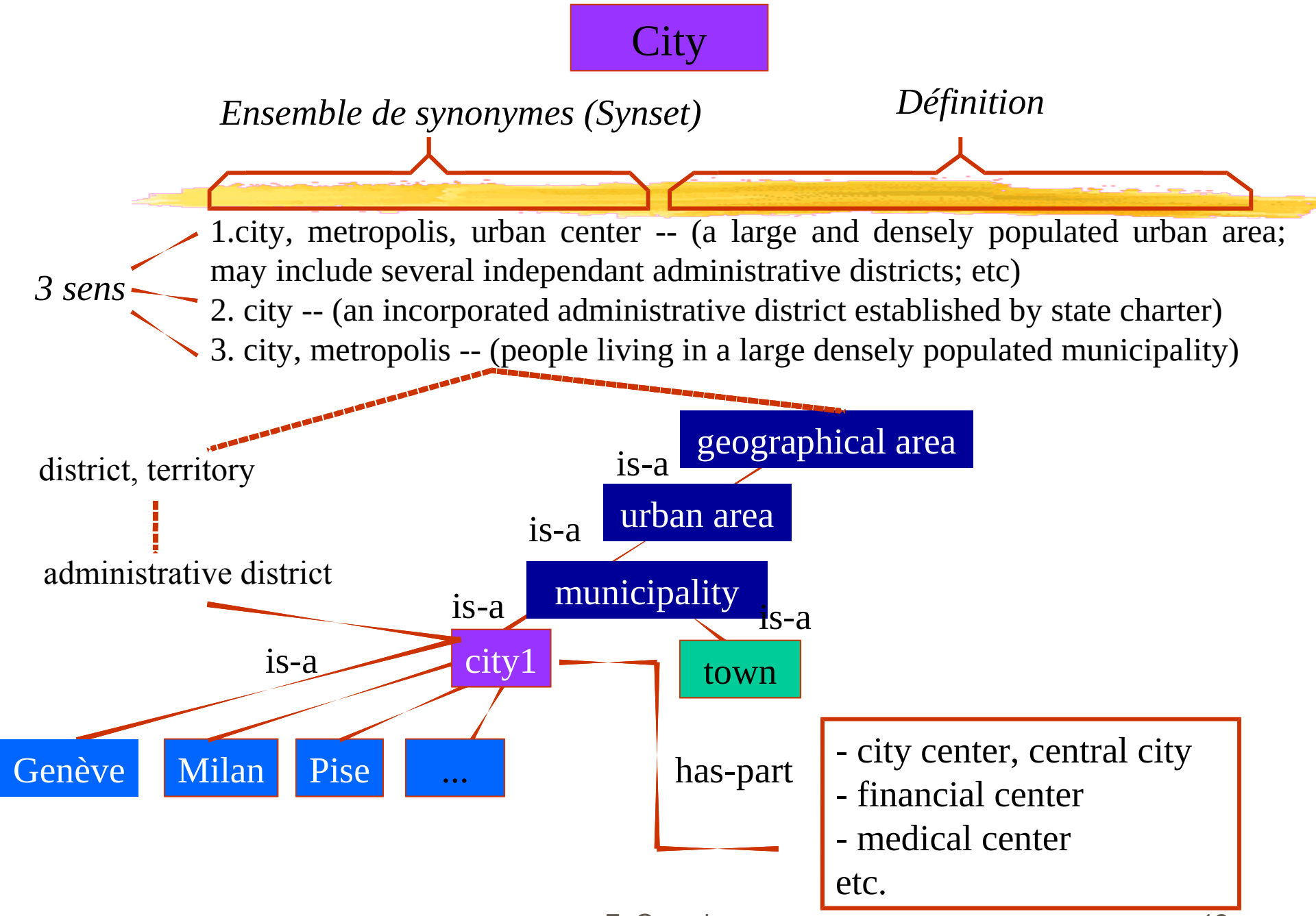
Les meronymes de building1

- Extrait des méronymes de *building1* :
  - {exterior door, outside door} (a doorway that allows entrance to or exit from a building)
  - Floor, level, storey, story (a structure consisting of a room or set of rooms at a single position along a vertical scale) "what level is the office on?"
  - foundation stone (a stone laid at a ceremony to mark the founding of a new building)
  - heating system, heating plant, heating, heat (utility to warm a building) "the heating system wasn't working"; "they have radiant heating"
  - interior door (a door that closes off rooms within a building)
  - Roof (a protective covering that covers or forms the top of a building)

Quelques hyperonymes de premier niveau de building

- *building3* : hyperonymes
  - {commercial enterprise, business enterprise, business}
  - the activity of providing goods and services involving financial and commercial and industrial aspects
  - "computers are now widely used in business"

- *building1* : hyperonymes
  - Structure, construction
  - a thing constructed; a complex entity constructed of many parts
  - *"the structure consisted of a series of arches"; "she wore her hair in an amazing construction of whirls and ribbons"*


City
Ensemble de synonymes (Synset)
Définition
3 sens
1.city, metropolis, urban center -- (a large and densely populated urban area; may include several independant administrative districts; etc)
2. city -- (an incorporated administrative district established by state charter)
3. city, metropolis -- (people living in a large densely populated municipality)
district, territory
administrative district
geographical area
is-a
urban area
is-a
municipality
is-a
is-a
city1
is-a
town
Genève
Milan
Pise
...
has-part
- city center, central city
- financial center
- medical center
etc.

Wordnet : un peu un lexique mais pas vraiment, un peu une ontologie mais pas complètement

- Lexique : information sur les mots
- Ontologie : organisation de concepts, indépendamment de la langue et des mots dans lesquels ces concepts s'incarnent

- WordNet organise les mots d'une langue (-> lexique) dans un réseau de relations, entre autres hiérarchiques (-> ontologie)

Wordnet : différences avec un lexique

- Division des mots suivant leurs catégories syntaxiques
- Pas de véritable définition du sens des mots :
  - synsets n'expliquent pas le concept sous-jacent mais permettent de le reconnaître
  - suppose que l'utilisateur connaît déjà les concepts

$\Rightarrow$ pas une théorie lexicale constructive mais différentielle

Wordnet : théorie différentielle du sens

- « Les ensembles de synonymes (synsets) n'expliquent pas ce que sont les concepts ; ils en posent l'existence. On suppose que les locuteurs anglophones ont déjà acquis ces concepts et sont en mesure de les reconnaître à partir des mots listés dans le synset. » (Miller et al., 1993, p. 5-6)
  - board1 (a committee having supervisory powers) *"the board has seven members"*
  - board2 (a flat piece of material designed for a special purpose) *"he nailed boards across the windows"*
  - *board6, table6 (food or meals in general) "she sets a fine table"; "room and board"*
  - dining table1, board9 (a table at which meals are served) *"he helped her clear the dining table"; "a feast was spread upon the board"*

Pas les mêmes relations pour les différentes catégories syntaxiques représentées

- Les quatre principales catégories syntaxiques (nom, verbe, adjectif, adverbe) sont représentées séparément car
  - hypothèse cognitive : représentées différemment en mémoire
  - pas les mêmes relations suivant les catégories syntaxiques
    - noms liés entre eux par hyponymie/hyperonymie alors que les verbes le sont par troponymie
    - V1 est un **troponyme** de V2 ~ V2-er signifie V1-er d'une certaine façon.
      - *boîter*, c'est *marcher* d'une certaine manière
      - *déguster*, c'est *manger* d'une certaine manière
- Version 1.6 :
  - 94.000 formes nominales
  - 10.000 formes verbales
  - 20.000 formes adjectivales
  - 4.500 formes adverbiales

Les noms dans Wordnet

- 114 648 noms pour 79689 nounsynsets
- WN divise les noms en plusieurs hiérarchies (12) : organism, entity, abstraction, psychol. feature, nat. phenomenon, activity, event, group, location, possession, shape, state
- Chaque hiérarchie correspond à un champ sémantique distinct mais elles ne sont pas mutuellement exclusives
- Le nombre maximum de niveaux à l'intérieur de ces hiérarchies est 12
  - shetland pony/horse/equid/odd-toe ungulate/herbivor/ mammal/mammal /vertebrate/animal/organism/entity

Hiérarchies nominales

- Héritage multiple possible :
  - piano → instrument de musique
    → meuble
- Souvent dû à une double vision de fonction / de structure
  - {ribbon, band}: strip of cloth → structure
  - {ribbon, band}: ornment → fonction
  - {cairn} : ensemble de cailloux → structure
  - {cairn} : marqueur → fonction

Méronymies entre synsets

- Wordnet code la relation de méronymie (partie/tout) entre les synsets de noms : principalement dans les hiérarchies pour body, artefact, quantity
- Les méronymes sont des attributs qui s'héritent par le biais de l'hyponymie
- Problème : placer les méronymes au bon niveau de la hiérarchie
  - roue (wheel) méronyme de véhicule luge (sled) hérite de véhicule
  - pourtant : une luge n'a pas de roue !
- Solution (bricolage) : introduction de concepts intermédiaires
  - véhicule à roues

Les 3 sortes de méronymie utilisées dans Wordnet

Composant/objet complet	*pédalier/bicyclette*
Membre /collection	*arbre/forêt*
Matière/objet	*métal/voiture*
Portion/masse	*flocon/neige*
Sous activité/ activité	*payer/acheter*
Zone/lieu	*oasis/désert*

Les verbes dans wordnet

- wordnet 2.0 11306 verbes pour 13508 verbsynset
- verbes (avec éventuellement les prépositions *look up*)

Relations structurant les verbes

- l'implication lexicale (lexical entailment) : V1 entails V2
  - *quelqu'un V1* logiquement implique *quelqu'un V2*
  - V1 ne peut pas être fait sans que V2 soit fait (ou l' a été)
    - *ronfler* implique *dormir*
    - *acheter* implique *payer*
- la troponomie
  - V1 troponyme de V2 : V1-er est V2-er d'une certaine manière
    - *crier* troponyme de *parler*
    - *courir* troponyme de *marcher*
    - *téléphoner* troponyme de *communiquer*
  - différentes reltions de manière suivant l'aspect sémantique qui est ajouté à V1 par rapport à V2 ?
    - intention, medium, intensité, vitesse, manière,…

Exercice : chercher des troponymes

- de *couper*

- de transfert de possession

Troponymie, implication et inclusion temporelle

- la troponomie est un cas particulier de l'implication
  - *X crie* implique *X parle*
- un troponyme est temporellement co-extensif de son sur "sur-verbe"
  - *crier* implique *parler*

  à chaque moment où X crie, X forcément parle
- ce n'est pas le cas pour un impliquant non-tropononyme
  - *acheter* implique *payer* et *acheter* non troponyme de *payer*
    - *payer* inclus dans *acheter*
  - *ronfler* implique *dormir* et *ronfler* non troponyme de *dormir*
    - *ronfler* inclus dans *dormir*

Catégories de verbes : 15 grandes familles

- ***Soin du corps et vitalité:*** verbs of grooming, dressing and bodily care
- ***Changement :*** change of size, temperature, intensity, etc.
- ***Cognition*** verbs of thinking, judging, analyzing, doubting, etc.
- ***Communication*** verbs of telling, asking, ordering, singing, etc.
- ***Competition*** verbs of fighting, athletic activities, etc.
- ***Consommation*** verbs of eating and drinking, using, ingesting
- ***Contact*** verbs of touching, hitting, tying, digging, etc.
- ***Creation*** verbs of sewing, baking, painting, performing, etc.
- ***Emotion*** verbs of feeling
- ***Mouvement*** verbs of walking, flying, swimming, etc.
- ***Perception*** verbs of seeing, hearing, feeling, etc.
- ***Possession*** verbs of buying, selling, owning, and transfer
- ***Interactions sociales*:** verbs of political and social activities and events
- ***Etats*** verbs of being, having, spatial relations
- ***Météorologie*** verbs of raining, snowing, thawing, thundering, etc.

Record

- Les sens2 and 1 de *change* a le plus de troponyms, puis {*be*}.

Les adjectifs dans Wordnet

- WN 1.5 contient 16 428 synsets d ’adjectifs (où des noms, des participes et des GPs sont inclus qui fonctionnent comme modificateurs e.g., home dans home cooking)

- WN divise les adjectifs en deux grandes catégories : les adjectifs descriptifs et les adjectifs relationnels
  - adjectifs descriptifs (*big, interesting,possible) :* la plus grande catégorie
  - adjectifs relationnels i.e., reliés (par dérivation) à un nom (*electrical, dental, fraternal, presidential, nuclear*) : classe beaucoup plus petite

Différence adjectifs descriptifs/ relationnels

- Un adjectif descriptif est un adjectif dont le rôle est d'assigner une valeur à un attribut d'un nom
  - le nom *paquet* a pour attribut poids dont la valeur peut être spécifiée par l'adjectif *lourd*
- Il existe une échelle de valeurs correspondant à ce nom, se traduisant par des adverbes *très, extrèmement,…*
  - *ce paquet très lourd* *cet homme assez intelligent*
- Un adjectif relationnel n'est pas gradable
  - **cette bombe très nucléaire* **cette installation assez électrique*
- Difficile de coordonner les 2 types d 'adjectifs :
  - *une installation électrique et très haute*
- Mais… il existe des adjectifs qui sont les 2 !!
  - *musical* relationnel (qui réfère à la musique) descriptif
  - *a musical instrument* *a musical child*

Wordnet : adjectifs descriptifs

- WN associe aux adjectifs descriptifs des pointeurs vers les noms
- Deux relations structurent la classe des adjectifs descriptifs : la similarité et l'antonymie

{damp, watery, moist, humid, soggy} ↔ wet

{parched, arid, anhydrous, sere, dried-up} ↔ dry

Wordnet : adjectifs relationnels

- reliés sémantiquement (et morphologiquement) à des noms
- fonction = classificateur e.g., musical instrument identifie un type d 'instrument
- ne réfèrent pas à un attribut du nom qu 'ils modifient
- n 'ont souvent pas d 'antonymes
- Wordnet :
  - regroupe les adjectifs relationnels (2 832 synsets)
  - les relie par des pointeurs aux noms dont ils sont dérivés
  - mais ne les structurent pas entre eux

Conclusion

- Avantages :
  - Une des rares ressources pour la langue générale (anglais) disponible en ligne
  - Organisation de la base lexicale assez riche grâce aux diverses relations représentées
  - Eurowordnet ressource multilingue

- Inconvénients :
  - Distinctions de sens très (trop) fines, sans méthodologie précise pour les découper, sans repérer des processus lexicaux type métonymie, métaphores…
  - Pas de définition des mots
  - Aucune information syntaxique, morphologique, de dérivation,…

Applications de wordnet

- utilisation de WordNet en recherche d'informations :
  - pour représenter les documents
  - pour étendre la requête de l'utilisateur (ajout de synonymes, par exemple, pour augmenter le rappel, c'est-à-dire la proportion de documents pertinents rapportés)
  - Acquisition de relations sémantiques
  - Désambiguïsation sémantique
- pour l'étiquetage sémantique de corpus
- pour la structuration et catégorisation des documents

WordNet comme ressource lexicale et composant pour NLP

- Ressource lexicale permet d'incorporer certaines connaissances lexico-sémantiques au traitement des textes
  - pour indexer les documents
  - pour réduire les différences de vocabulaire entre les textes, entre les textes et les questions sur ces textes

- Applications
  - pour la recherche d'informations
  - pour l'extraction d'informations
  - pour les systèmes de questions/réponses

⇒ Enrichissement de la représentation avec des synonymes, hyperonymes, ...

Dans les systèmes de question-réponse

- M. Papa, S. Harabagiu (Dallas)
  - pour le traitement de la question : aider à déterminer le type attendu de la réponse
    - *What flowers did Van Gogh paint ?* → type ‘flower1’
  - pour l’expansion de requêtes (enrichir les mots-clés extraits de la question)
    - flower → hyponyms de flower1 : orchids, petunias
    - paint → synonym de paint : draw,...
      - *Van Gogh draws orchids, petunias,..*
    - *When was Brandeburg door built ?*
    - *The Brandeburg door construction…* → build1 synonym de erect1

Différentes sources de ce cours

- Agnès Tutin (Grenoble)
- Claire Gardent (Nancy)
- Jacques Moeschler (Genève)
- Benoit Habert (Nanterre)
- ...