

Analyse et représentation de la langue

Damien Nouvel



Plan

1. Analyses lexicales et morphologiques
2. Analyses syntaxiques
3. Vers la sémantique

Correction orthographique

- ▶ **Recherche dans une base lexicale**
 - Liste des mots pour une langue donnée
 - Comparaison mots du texte vs lexique
 - Disposer de toutes les formes possibles : **couverture**

⇒ Complexe : par exemple $1000 * 50K = 50M \dots$
- ▶ Optimisation de la recherche
 - Utilisation d'automates
 - Arbre des préfixes (TRIE)
- ▶ Besoin d'informations contextuelles

Couverture du lexique

- ▶ Difficultés de couverture
 - Langues « agglutinantes » (morphologie / composition)
 - Morphologie, dont **flexions** (déclinaisons / conjugaisons)
 - ⇒ **Paradigmes** d'agglutination et de flexion
 - Variantes d'écritures
 - Noms propres (classe **ouverte**) acronymes, etc.
 - ⇒ Automates de reconnaissances partielles

Applications

- ▶ Éditeurs de texte
 - ▶ Saisie prédictive / autocorrect (T9)
 - Handicap (aveugle, moteurs)
 - Grand public
- ⇒ Logiciels **hors-ligne** (contrairement aux reconnaissances)
- ▶ Détection de nouveaux mots (néologismes, noms propres, etc.)

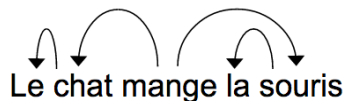
Plan

1. Analyses lexicales et morphologiques
2. Analyses syntaxiques
3. Vers la sémantique

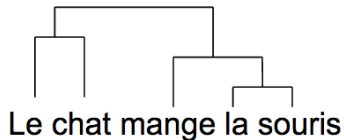
Lier les mots dans les énoncés

- ▶ Niveau de représentation
 - Exploite les approches morphologiques / lexicales
 - ⇒ Importance des **catégories morpho-syntaxiques** (POS)
 - Détermine les liens / relations / **dépendances** entre les mots
 - Identifie des **syntagmes**
- ▶ Modélisation de la « grammaire » (SOV, VSO, etc.)
- ▶ Deux catégories d'approches
 - **Constituants** : arbres syntaxiques (projectivité)
 - **Dépendances** : relations entre mots (Tesnière, 1940)
 - ⇒ Possibilité de convertir de l'un à l'autre
 - ⇒ Approches : LFG, GPSG, TAG, HPSG ...

Exemple



(Bourigault / SYNTEX)



Quelques difficultés

- ▶ Le rattachement prépositionnel
 - « Il voit le chat de sa sœur »
 - « Il voit le chat de sa fenêtre »
- ▶ Triple ambiguïté locale
 - « Il voit l'homme dans le parc avec des jumelles »
- ▶ Les propositions (relatives, subordonnées, etc.)
- ▶ Quelle représentation pour la coordination
- ▶ Traitement des anaphores (syntaxe et sémantique)

Démo

- ▶ Test avec **Stanford Core NLP** (Manning)
 - Analyseur : <http://corenlp.run>
 - Documentation : <https://stanfordnlp.github.io/CoreNLP/>
- ▶ Quelques essais avec **FRMG** (de la Clergerie)
 - Analyseur : http://alpage.inria.fr/frmgwiki/frmg_main/frmg_server
 - Portail : <http://alpage.inria.fr/frmgwiki/>

Plan

1. Analyses lexicales et morphologiques
2. Analyses syntaxiques
3. Vers la sémantique

Profondeur de la syntaxe

- ▶ Analyse syntaxique
 - Morpho-syntaxe
 - Syntaxe de surface / parenthésage (syntagmes)
 - Syntaxe « grammaticale »
- ▶ Syntaxe « profonde »
 - Sujet du verbe *vs* agent de l'action (passif)
 - Objet : patient, récepteur, etc.
 - Sujet distant (anaphores / coréférences ...)

⇒ Interface syntaxe / sémantique

Les entités nommées

Tous Maps Actualités Images Vidéos Plus Paramètres Outils

Environ 1 150 000 résultats (0,89 secondes)

Inalco | Institut National des Langues et Civilisations Orientales
www.inalco.fr/ ▼
 Établissement public d'enseignement supérieur et de recherche, l'Inalco enseigne et mène des recherches sur les langues d'Europe centrale et orientale, ...

<p>Formations Licences - Formations et diplômes - Masters - Apprendre une langue</p>	<p>Formations et diplômes Licences - Masters - Passeport Langues O - ...</p>
<p>Langues et civilisations Coréen - Arabe littéral - Persan - ...</p>	<p>Emplois du temps et examens Emplois du temps - Calendrier universitaire - Examens - ...</p>
<p>S'inscrire à l'Inalco Je m'inscris en master à l'Inalco ... Comment s'inscrire en master?</p>	<p>L'Institut L'Institut. Une centaine de langues et civilisations sont enseignées ...</p>

Institut national des langues et civilisations orientales — Wikipédia
https://fr.wikipedia.org/wiki/Institut_national_des_langues_et_civilisations_orientales ▼
 L'Institut national des langues et civilisations orientales (INALCO), dit Langues O (prononcer Langzo), est un établissement français d'enseignement supérieur ...
[Histoire](#) · [Organisation](#) · [Les équipes de recherche](#) · [Les Presses de l'Inalco](#)

Institut national des langues et civilisations orientales (INALCO ...
<https://www.sorbonne.fr/.../institut-national-des-langues-et-civilisations-orientales-inal...> ▼

inalco
 Institut national des langues et civilisations orientales
 Voir les photos
 Extérieur

Institut national des langues et civilisations orientales ★
 Site Web Itinéraire
 Etablissement d'enseignement supérieur à Paris, France

L'Institut national des langues et civilisations orientales, dit Langues O', est un établissement français d'enseignement supérieur et de recherche chargé d'enseigner les langues et civilisations autres que celles originaires d'Europe occidentale. [Wikipédia](#)

Adresse : 65 Rue des Grands Moulins, 75013 Paris
Téléphone : 01 81 70 10 00
Création : 30 mars 1795
Président : Manuelle Franck.
Enseignants-chercheurs : 319 (245 titulaires)
Type : Grand établissement (EPSCP)
Nombre d'inscrits : 9 188 (2007)

Les entités nommées

- ▶ Pas de définition stable
 - « Noms propres et quantités d'intérêt » (MUC)
 - « Entités du monde concret qui ont un nom » (ESTER)
 - « Expression qui réfère à une entité unique » (Ehrmann)
 - « Objets mentaux pour la logique » (Nouvel)
 - ...
- ▶ Deux grandes catégories linguistiques
 - **Noms propres**
 - ⇒ Limites : « Superman », « l'Histoire », etc.
 - **Descriptions définies** (syntagmes nominaux définis)
 - ⇒ Limites : « cette voiture », « le roi des forains », « l'occident »
- ⇒ Tenir compte des **synonymes**, **homonymes**, **métonymies**
- ⇒ Beaucoup d'applications ...prétraitement ?

Structurer les textes en connaissances

▶ Des **formats** plus ou moins **structurés**

- Texte brut : sauts de lignes
- PDF : graphiquement stable, mais peu structuré
- HTML : sections, styles, liens, micro-formats
- Wikipedia : articles et info-boxes
- BabelNet : <http://babelnet.org/synset?word=Inalco&lang=FR>

⇒ Structuration : du texte aux **graphes** (valués)

▶ Quelques **réseaux sémantiques**

- WordNet
- Graphes conceptuels
- Web sémantique (RDF/OWL)
- JeuxDeMots <http://www.jeuxdemots.org>

⇒ Interactions **homme-machine** et **machine-machine**

Plongements de mots

- ▶ Sémantique distributionnelle
 - « You shall know a word by the company it keeps » (Firth, 57)
 - Analyse « distributionnelle » des mots en contexte (Harris, 60)
 - ▶ Mélange de deux axes d'analyse
 - **Syntagmatique** : les successions de mots
 - **Paradigmatique** : les substitutions de mots
 - ▶ Méthode non-supervisée de positionnement (sémantique)
 - Collection de corpus de texte **très volumineux**
 - Choix de **paramètres** (dimensions)
 - Recherche de **correlations entre mots et leurs contextes**
- ⇒ **Embeddings** (Collobert, 2011), **Word2Vec** (Mikolov, 2013), FastText (Bojanowski, 2016)
- ⇒ Démo : <http://embeddings.sketchengine.co.uk>