

Les différentes technologies de la langue

Pierre Zweigenbaum / Joseph Mariani

LIMSI, CNRS, Université Paris-Saclay

11 mai 2016

Technologies de la langue : composants et ressources

Technologies de la langue

Composants et ressources

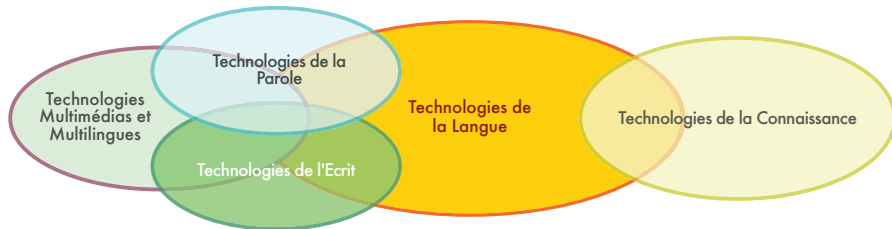
Applications de traitement de la langue

Textes, production, traduction

Langue parlée, écrite, signée

Textes, information, connaissance

Technologies de la langue



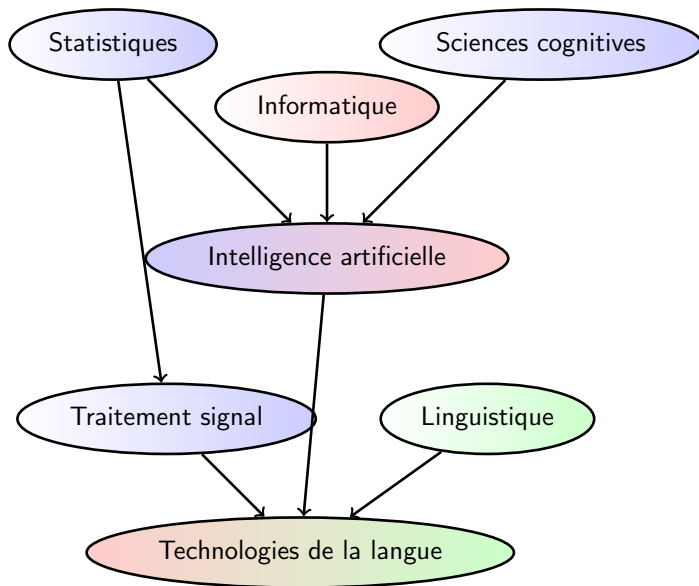
- ▶ Technologies de l'information qui sont spécialisées pour traiter le langage humain

langue parlée

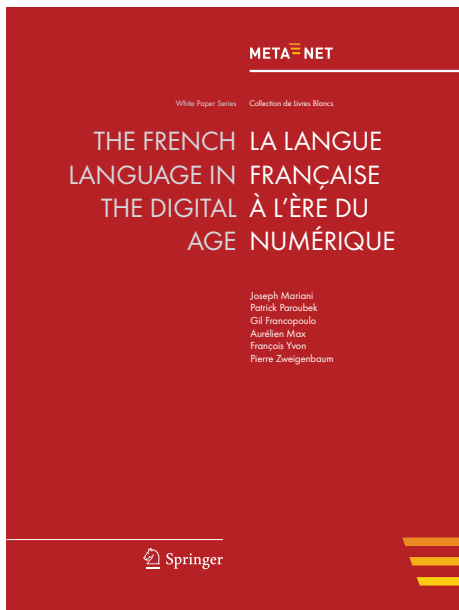
langue écrite

langue des signes

Disciplines mobilisées



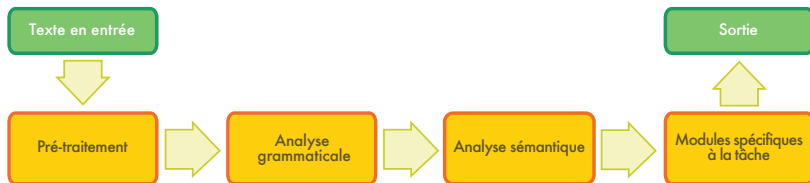
Mariani *et al.* (2012)
Livre blanc META-NET



Composants de traitement de la langue

Composants classiques dans une architecture simplifiée

- ▶ Traitements préalables
 - ▶ Nettoyage, suppression du formatage, détection de la langue
- ▶ Analyse grammaticale
 - ▶ Détection des phrases, des mots ; des noms, verbes, etc. ; des sujets, compléments, etc. : structure de chaque phrase
- ▶ Analyse sémantique
 - ▶ Désambiguïsation des mots selon le contexte (*avocat*), résolution des références (*elle, la voiture*), représentation du sens d'une manière lisible par la machine
- ▶ *+ Composants spécifiques à la tâche visée*



Composants de traitement de la langue

État des lieux pour le français

	Quant	Disp	Qual	Couv	Matur	Péren	Adapt
Analyse morpho-lexicale et morpho-syntaxique	4	4	4	4	4	3	3
Analyse syntaxique	4	4	4	4	4	2	2
Sémantique des phrases	2	2	2	1	2	1	2
Sémantique des textes	2	1	3	2	2	2	1
Traitement avancé du dis- cours	2	2	2	2	2	2	1

Ressources pour le traitement de la langue

Les technologies de la langue ont besoin de « ressources » pour fonctionner

	Quant	Disp	Qual	Couv	Matur	Péren	Adapt
Lexiques, bases terminologiques	4	3	4	3	4	4	3
Grammaires	3	2	3	3	3	2	2
Thésaurus, WordNets	3	3	2	1	3	3	3
Ressources ontologiques pour la connaissance du monde (<i>upper models</i> , données liées)	2	1	2	1	2	1	1
Corpus de référence	3	1	4	3	4	4	3
Corpus syntaxiques (corpus arborés)	4	4	3	3	3	3	2
Corpus sémantiques	2	2	2	1	2	2	2
Corpus de discours	1	2	2	2	1	1	1

Technologies de la langue : composants et ressources

Technologies de la langue

Composants et ressources

Applications de traitement de la langue

Textes, production, traduction

Langue parlée, écrite, signée

Textes, information, connaissance

Correction orthographique

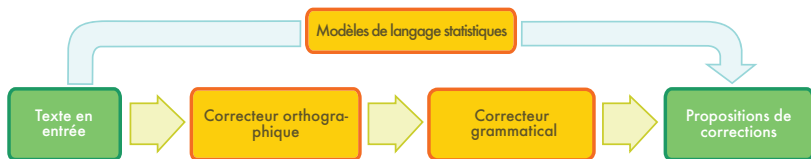
Aides à la rédaction

Correcteur d'orthographe et grammaire français

Je ne crois pas me tromper.

Employez ici l'infinitif **tromper** au lieu de la forme verbale **trompé**.

Correcteur Cordial, site Reverso



Traduction automatique

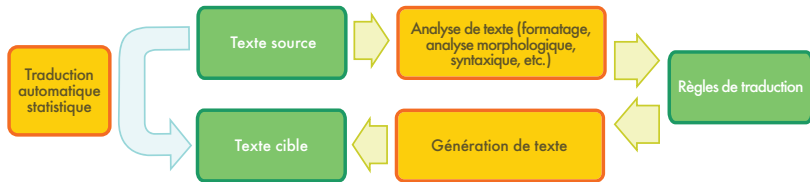
Voir l'exposé de François Yvon

Traduire cette description en Français

Our comfortable private room (with fridge) has a true city feeling! It comfortably fits for two persons. It's centrally located on a quiet area of the Piran, just a few steps from the beach, trade-market, tobacco shop, restaurants, bars, churches.

* Notre chambre privée confortable (avec frigo) a un vrai sentiment de la ville! Il adapte confortablement pour deux personnes. Il s'agit d'un lieu tranquille du Piran, à seulement quelques pas de la plage, le commerce sur le marché, bureau de tabac, restaurants, bars, églises.(URL

Traduction Google sur Airbnb



Qnt

Dsp

Qal

Cvt

Mtr

Prn

Adp

Résumé automatique, Génération de texte

D'autres aides à la rédaction

- ▶ Résumé par extraction : sélection des phrases clés du document

Qnt

Dsp

Qal

Cvt

Mtr

Prn

Adp

- ▶ Génération de texte : création de nouvelles phrases, de nouveaux textes
 - ▶ bulletin météo, comptes rendus médicaux

Qnt

Dsp

Qal

Cvt

Mtr

Prn

Adp

Technologies de la langue : composants et ressources

Technologies de la langue

Composants et ressources

Applications de traitement de la langue

Textes, production, traduction

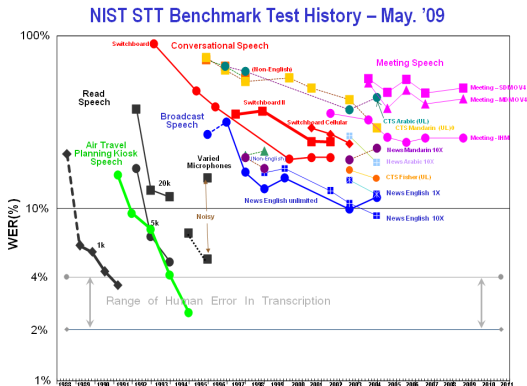
Langue parlée, écrite, signée

Textes, information, connaissance

Reconnaissance de la parole

Parole → texte

- ▶ Commande vocale
- ▶ Dictée automatique
- ▶ Transcription de parole : nouvelles radiotélévisées, réunions
- ▶ ...
- ▶ Détection d'émotions dans la parole



Qnt

Dsp

Qal

Cvt

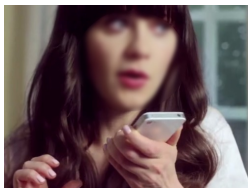
Mtr

Prn

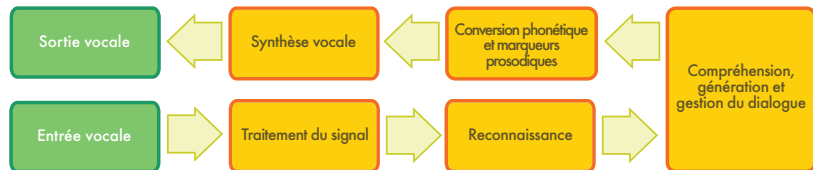
Adp

Dialogue oral

Oral ↔ Oral



Siri d'Apple



Qnt

Dsp

Qal

Cvt

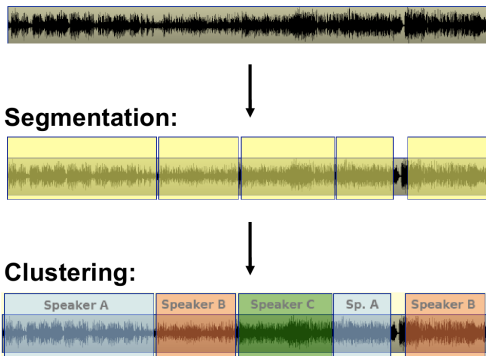
Mtr

Prn

Adp

Détection du locuteur (qui parle)

Segmentation et regroupement des locuteurs dans un flux audio, dans un flux vidéo

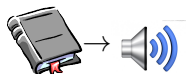


SCALE speaker diarization tutorial, 01/13/2010, Christian Müller

Indexation des segments, accès direct à ce qu'a dit un locuteur

Synthèse vocale

Texte → parole



- ▶ Lecture à voix haute
 - ▶ Synthèse d'annonces (train, gares, etc.)
 - ▶ Lecture de réponses

Qnt

Dsp

Qal

Cvt

Mtr

Prn

Adp

Traitement de l'écrit

Image → texte

- ▶ Numérisation de textes
- ▶ Reconnaissance de caractères



MODERN war means mechanized armies, tanks, aeroplanes-thousands of them-and guns, in addition to men and munitions, and to equip and maintain a modern army for defending our homes will cost very large sums of money. Every patriot Indian can help to make India strong and thus protect his home and family by saving as much as possible and lending his savings to Government. The Government of India have issued Defence Loans to suit the convenience of different classes of people. The object of these Defence Loans is to raise money to expand and equip our army, navy and air force for the defence of our country. By subscribing to these loans you will be paying the best insurance for your freedom and happiness.

Free OCR To Word

Traitement de la langue des signes

Une autre langue

- ▶ Langue visuo-gestuelle
- ▶ Modélisation de la langue des signes
- ▶ Synthèse de textes en LSF
- ▶ Analyse d'énoncés



Gare SNCF, Websourd/LIMSI

Technologies de la langue : composants et ressources

Technologies de la langue

Composants et ressources

Applications de traitement de la langue

Textes, production, traduction

Langue parlée, écrite, signée

Textes, information, connaissance

Recherche sur le web

Recherche dans des bases de documents, dans des systèmes d'information

découverte neptune



Pages Web



Pré-traitement

Traitement sémantique

Indexation

Mise en correspondance et calcul de la pertinence

Pré-traitement

Analyse de la demande

Demande de l'utilisateur

Résultat de la recherche

Qnt

Dsp

Qal

Cvt

Mtr

Prn

Adp

Découverte de Neptune — Wikipédia

[Éditer](#) | [Ajouter une référence](#) | [Découverte de Neptune](#) •

La découverte de Neptune est la première découverte d'un objet céleste majeur grâce au calcul mathématique avant de l'être par l'observation. Langages…

Première observation - Intégration dans l'article d... - Cours et la découverte

^[P] La découverte de Neptune par Le Verrier (1846) - BibNum

<https://www.bibnum.education.fr/bases/bnf/bnf-neptune-analyse.pdf> •

La grande Uranie a été découverte en 1781 par William Herschel (1738... de la Terre, le premier calculer pour la date de la découverte (23 septembre 1846).

La découverte de Neptune - JeSuisCultive.com

www.jesuiscultive.com/tp/tparticle101 •

JEUSUISCULTIVE.COM - articles de culture générale, quiz, divertissements, anecdotes et réponses à des

questions et à des quiz de base. La découverte de Neptune.

DÉCOUVERTE DE NEPTUNE - Encyclopædia Universalis

<https://www.universalis.fr/encyclopedie/d%C3%A9couverte-de-neptune/> •

DÉCOUVERTE DE NEPTUNE : le jour même de la réception de la lettre de Le Verrier, le 23 septembre 1846, la nouvelle planète, qui sera nommée Neptune.

^[P] La découverte de neptune, ou vive les maths! - planetastronomy.com

www.planetastronomy.com/fr/comptes-rendus/neptune.pdf •

La découverte de Neptune par Jean-Pierre... Dans les années suivant la découverte d'Uranus par

découverte une nouvelle planète - Neptune. C'est cette...

Recherche de réponses à des questions

Réponse précise plutôt que document entier

Qui a découvert Neptune ?

→

Urbain Le Verrier

- ▶ Analyse de la question
- ▶ Recherche de réponses
 - ▶ Dans des bases de données / de connaissances
 - ▶ Dans des bases de questions et réponses (FAQ)
 - ▶ Dans des textes
 - ▶ Mixte

Qnt

Dsp

Qal

Cvt

Mtr

Prn

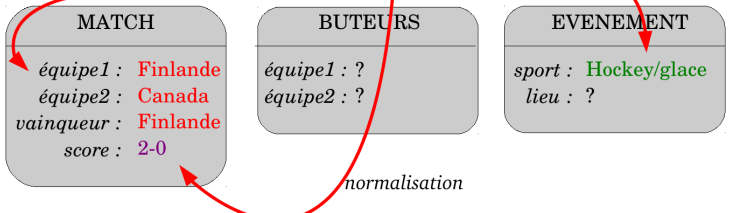
Adp

Extraction d'information

Inclut la détection d'entités, de relations, de négations, d'opinions, de sentiments, la fouille de textes, l'anonymisation, etc.

La Finlande a battu le Canada 2 à 0 lors d'un match du premier tour du tournoi masculin de hockey sur glace.

- Résultat :



Transparent : Xavier Tannier

- Nombreuses applications

Qnt

Dsp

Qal

Cvt

Mtr

Prn

Adp

Classification de texte

Tâche générique avec de nombreuses applications

- ▶ Ce message est-il un pourriel ?
- ▶ Quel est le thème de cet article ?
- ▶ Qui est l'auteur de ce livre ?
- ▶ Dans quelle catégorie faut-il indexer ce brevet / cet article / cette recette ?
- ▶ En quelle langue est écrit ce message ?
- ▶ Quel est le degré de lisibilité de cette brochure ?

En conclusion

	Quant	Disp	Qual	Couv	Matur	Péren	Adapt
Traduction automatique	5	4	4	3	4	3	3
Génération de langue	2	1	2	2	2	1	2
Reconnaissance de la parole, dialogue oral	4	3	4	4	4	3	3
Synthèse vocale	4	3	4	4	4	3	3
Gestion de dialogue	3	2	3	3	3	2	2
Recherche d'Informations	4	5	5	4	5	4	4
Résumé automatique, Réponse aux questions, Technologies avancées d'accès à l'information	3	3	3	3	3	2	2
Extraction d'Informations	3	3	4	3	4	3	3