Un rapide survol historique

Traitement Automatique des Langues

- Tentative de reproduire avec des ordinateurs certains traitements linguistiques réalisés par l'humain
- On n'essaye pas de reproduire le fonctionnement du cerveau humain
- On essaye de tirer le meilleur parti de l'ordinateur
- Les modèles sont conçus pour minimiser les erreurs sur la tâche qu'ils doivent résoudre
- Exemples
  - Compréhension de la parole
  - Traduction automatique
- Trois grandes étapes
  1. Modèles linguistique
  2. Modèles statistiques
  3. Modèles end-to-end

Modèles linguistiques

- Un partenaire naturel pour le TAL : la grammaire générative
  La grammaire d'une langue propose d'être une description de la compétence intrinsèque du **locuteur-auditeur idéal**. *Si la grammaire est, de plus,* **parfaitement explicite** *(...), nous pouvons (...) l'appeler grammaire générative.*
  – Noam Chomsky *Aspects de la théorie syntaxique, 1971*
- Naissance de la **linguistique computationnelle**.
- Un programme scientifique clair :
  réaliser le locuteur-auditeur idéal à l'aide de l'ordinateur.

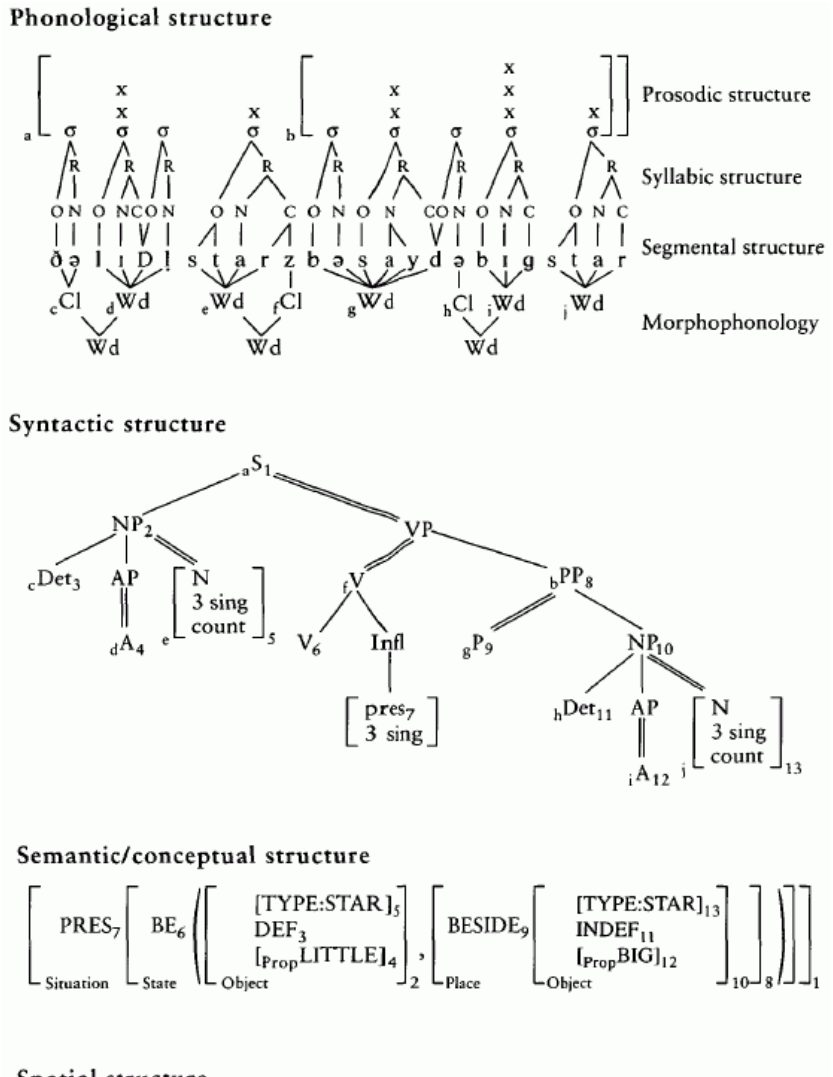
Symboles, Relations, Règles et Structures

- **Symboles** représentant des unités abstraites, intelligibles
  Phonèmes, Syllabes, Morphèmes, Parties de discours, Syntagmes . . .
- **Relations** entre symboles
  Dominance, précédence, dépendance . . .
- **Règles** de manipulation de symboles
  Règles de génération, Règles de transduction, Transformations, Mouvements, . . .
- **Structures**
  Phonétique, Phonologique, Morphologique, Syntaxique, Sémantique, Discursive . . .

Une forêt de symboles

Ray Jackendoff – Foundations of Language, 2002

Fig. 1.1 Structure of *The little star's beside a big star*



Le locuteur-auditeur est idéal mais le monde ne l'est pas !

- Avec Fillon et une France qui donne l'impression de se **filloniser** ou de se **macroniser** de plus en plus, pas si tôt...
  www.latribune.fr, 1er décembre 2016
- bonjour **voila** depuis samedi **j** ai **changer** toute mon **instalation** **live box** plus **decodeur** mais quand on regarde la tv **sa s arrete** souvent et apres **sa** repart
  Données tchat Orange
- Ma grand-mère chez elle le salon c'est de la moquette
  José Deulofeu - professeur de linguistique AMU

In theory there is no difference between theory and practice.
In practice there is.
Yogi Berra

Modèles statistiques

- D'une part des modèles linguistiques de description de la langue
- D'autre part des données naturelles décrivant la manière dont les humains se servent de la langue
- Comment utiliser ces données disponibles en grande quantité ?
  - Associer des probabilités aux règles et aux structures.
  - Détecter des régularités pour apprendre de nouvelles règles.

Modèles statistiques

- Données : (X, Y)
  - X = données observables (texte, signal de parole)
  - Y = données en général non observables, structure linguistique (parties de discours, morphèmes, arbres syntaxiques ...)
- Les modèles permettent de calculer $P(Y|X)$
- et de choisir une solution optimale $\hat{Y} = \arg\max_{Y \in \mathcal{Y}} P(Y|X)$

Apprentissage de représentations

- Les symboles et les stuctures hérités de la linguistiques sont ils optimaux pour le TAL ?
- Ils ont été postulés par l'humain pour décrire la langue et communiquer avec d'autres humains
- Ne vaut il pas mieux laisser à l'ordinateur la tâche de découvrir les représentations les plus adéquates pour une tâche particulière ?

Représentations continues des mots

- Un mot n'est plus représenté par un symbole mais par un vecteur dans $\mathbb{R}^n$
- La distance entre deux mots dans $\mathbb{R}^n$ reflète la "proximité" de ces derniers

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

Plus proches voisins (Collobert et al. 2011)

You shall know a word by the company it keeps – Firth 1957

- Sémantique distributionnelle
- Deux mots sont d'autant plus proches qu'ils partagent de contextes communs.
- On crée une tâche de classification dont le but est de discriminer séquences correctes et incorrectes :
  - mobilisation contre **un** cinquième mandat $\rightarrow 1$
  - mobilisation contre **commerces** cinquième mandat $\rightarrow 0$
- L'ordinateur apprend une **représentation** des mots optimale pour cette tâche.
- La méthode est généralisable à tout type de symbole.
- Mais on garde les catégories linguistiques.

Modèles *end to end*

- Tâche complexe, pour laquelle on dispose d'entrées et de sorties, tel que la traduction automatique.
- On fournit à l'ordinateur des couples de phrases traductions l'une de l'autre.
- L'ordinateur apprend un couple encodeur-décodeur :
  - L'encodeur apprend une représentation de la phrase source dans $\mathbb{R}^n$.
  - Le décodeur génère depuis la représentation une phrase dans la langue cible.
- Il n'y a plus de représentation explicite de la structure de la phrase !

Linguistique continue

- Symboles, règles et structures sont remplacés par des vecteurs, des matrices et quelques fonctions non linéaires.
- Les modèles sont devenus inintelligibles :
  A l'exception de l'entrée et de la sortie, les représentations intermédiaires ne sont pas compréhensibles.
- On arrive à reproduire sans toujours comprendre.
- Quels liens avec le traitement de l'information par le cerveau ?